DOCUMENT RESUME

ED 318 310 HE 023 298

AUTHOR	Adelman, Clifford, Ed.
TITLE	Signs & Traces: Model Indicators of College Student Learning in the Disciplines.
INSTITUTION	Office of Educational Research and Improvement (ED), Washington, DC.
REPORT NO	OR-89-538
PUB DATE	Sep 89
NOTE	195p.
AVAILABLE FROM	Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	Biology; Chemistry; Computer Science; Engineering; *Evaluation Methods; Higher Education; Majors (Students); Models; *Outcomes of Education; Physics; *Student Evaluation; Undergraduate Study
IDENTIFIERS	*Indicators

ABSTRACT

This report examines possible national indicators for learning outcomes in the individual disciplines of higher education. It consists of versions of five project final reports, each underscoring a distinct approach to developing a model in the context of a specific discipline. Each model applies, however, to several similar disciplines. Stressed in all the models are creative approaches to student assessment which include both criterion-referenced and norm-referenced information. Papers have the following titles and authors: "Models for Developing Computer-Based Indicators of College Student Learning in Computer Science" (Jerilee Grandy); "A Model for Assessing Undergraduate Learning in Mechanical Engineering" (Jonathan Warren); "Model Indicators of Student Learning in Undergraduate Biology" (Gary Peterson and Patricia Hayward); "A Study of Indicators of College Student Learning in Physics" (James Terwilliger, J. Woods Halley, and Patricia Heller); "Model Indicators of Undergraduate Learning in Chemistry" (George Bodner). Papers are followed by references, and appendixes, which detail model recommendations and criteria. (DB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED318310

HE 023 298

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

HE

Signs & Traces

Model Indicators of College Student Learning in the Disciplines

Clifford Adelman
Office of Research, U.S. Department of Education
George M. Bodner, Purdue University
Jerilee Grandy, Educational Testing Service
Gary Peterson and Patricia Hayward,
The Florida State University
James S. Terwilliger, J. Woods Halley and Patricia Heller,
The University of Minnesota
Jonathan Warren, Research in Higher Learning

Edited by Clifford Adelman

Office of Research
September 1989

U.S. Department of Education
Lauro F. Cavazos
Secretary

Office of Educational Research and Improvement
Bruno V. Manno
Acting Assistant Secretary

Office of Research
Milton Goldberg
Director

Information Services
Sharon K. Horn
Acting Director

For sale by the Superintendent of Documents, U.S. Government Printing Office
Washington, D.C. 20402

SN 065-000-00384-8

Acknowledgments

The very existence of the project that resulted in this volume is due to many in the Office of Educational Research and Improvement, but we owe most to the support, prodding, and counsel of Emerson Elliott, Acting Commissioner of the National Center for Education Statistics.

For the particular task of shepherding this publication through the bureaucracy and shaping it so that it is editorially and visually presentable to the general public, I am indebted to my colleagues Joseph Conaty, Charlene Medley, Lance Ferderer, Kate Dorrell, Margery Martin, and Philip Carr.

For their voluntary service in reviewing the text of the original Request for Proposals (RFP), the proposals themselves, and interim or final reports from the projects we sponsored under the RFP, we are grateful to the following individuals, listed with their affiliations at the time they served:

Richard Berry, National Science Foundation
David Breneman, President, Kalamazoo College
Tom Carroll, Fund for the Improvement of Postsecondary Education
John A. Centra, Department of Higher Education, Syracuse University
Louis W. Goodman, Dean, School for International Studies, The American University
Richard Light, J.F.K. School of Government, Harvard University
Bruno V. Manno, Office of Educational Research and Improvement
Jerry R. Mohrig, Department of Chemistry, Carleton College
Theodore Reed, National Science Foundation
Richard Schwartz, Dean of the Graduate School, Georgetown University
Irving Spitzberg, Association of American Colleges
Joyce D. Stern, National Center for Education Statistics
Jon Westling, Provost, Boston University
Alexandra Wigdor, National Academy of Sciences

Contents

Introduction: Indicators and Their Discontents	Clifford Adelman	1
Models for Developing Computer-Based Indicators of College Student Learning in Computer Science	Jerilee Grandy	11
A Model for Assessing Undergraduate Learning in Mechanical Engineering	Jonathan Warren	65
Model Indicators of Student Learning in Undergraduate Biology	Gary Peterson and Patricia Hayward	93
A Study of Indicators of College Student Learning in Physics	James S. Terwilliger, J. Woods Halley, and Patricia Heller	123
Model Indicators of Undergraduate Learning in Chemistry	George M. Bodner	149

Introduction:

Indicators and Their Discontents

Clifford Adelman
Office of Research, U.S. Department of Education

Signs and Traces is the second volume in a series that grew from the national discussion of recommendations made in Involvement in Learning: Realizing the Potential of American Higher Education. Sponsored and issued by the U.S. Department of Education in 1984, Involvement was the first of the current wave of reports on the status of higher education in the United States, and one that paid considerable attention to the question of how we know what college students learn between matriculation and graduation.

Four Contexts for a Question

This deceptively simple question spawned four related projects, each of which took the question into a different context where it assumed a life of its own. The four contexts called for distinct modes of sponsorship, and the reader might find it helpful to understand both the distinctions and the relationships in order to judge and use this volume.

The first version of the question focused wholly on assessment methods in a specifically higher education context. The recommendations in Involvement encouraged a tide of third-party assessment that was already rising in at least the public sector of higher education. Under legislative regulations, State Boards of Higher Education mandates, and other similar external pressures, thousands of college faculty and administrators newly involved in assessment needed a basic reference work and technical guide to the tasks they faced. Since we were responsible for drawing more attention to assessment in higher education, we were morally obligated to provide such a work. For that task, we determined a set of topics for which the existing literature did not provide adequate information on technical aspects of assessment in higher education (e.g., problems of construct validity in questionnaires, problems of reliability in performance assessments, difficulty levels in general education examinations, etc.) and contracted with a team of scholars who were expert in various testing and measurement issues for essays on those topics. The result was Performance and Judgement: Essays on Principles and Practices in the Assessment of College Student Learning, published in 1988.

In the second version of the question, the assessment of college student learning was subordinated to the curricular experiences of college students. In its simplest terms, the question became, "Do You Learn What You Study?" An obvious question, but one for which answers in higher education could not be found. To answer this question, the Office of Research conducted a contract competition under the title, "The Effects of Differential Coursework on Student Learning in College." This was a major 3-year undertaking by the winning team at Iowa State University, the results of which will be published in 1990. The project involved a sample of 1500 seniors in five 4-year colleges (each of a different type), and an analysis of their transcripts in relation to residual scores (that is, the difference between the predicted score and the actual score) on nine item-types in the Graduate Record General Examination.

What we will find out from this study is what patterns of coursework are most likely to improve a specific general learned ability (e.g., analytical reasoning) in different types of 4-year colleges. We will also find out how student learning styles affect the choice of coursework itself, and how to do course-cluster analysis of transcripts in order to provide better information to academic advisors and institutional researchers.

In the third version of the question, we focused on the relationship between the different forms and timing of students' college careers, along with coursework patterns, and their post-college experiences through early adulthood. Assessment in the sense of specific learning outcomes figures but indirectly in this analysis. To investigate the myriad of topics that arise under this rubric, staff in the Office of Research conducted intramural research using the survey and postsecondary transcript data from the National Longitudinal Study of the High School Class of 1972. This collection of data is the richest archive ever assembled on a generation of Americans, and follows them from age 18 to age 32 with significant attention to both the details of their college transcripts (whether or not they earned any degrees) and their subsequent experience in the labor market. The mode of intramural research was chosen because the extant data tapes required a significant amount of cleaning and correction, and OERI staff could handle those tasks more efficiently than others. Our work on this project was sponsored, in part, by the National Science Foundation and the U.S. Department of Labor, and we will be publishing a selection of these studies in late 1989 or early 1990.

The fourth context for the question is that of this volume. Having drawn attention to the importance of the learning outcomes of higher education, the question was raised as to whether national indicators of those outcomes could be constructed. It was agreed at the outset that the most credible contexts for such indicators would lie in the individual disciplines and fields, where consensus concerning the outcomes of undergraduate education was likely to be higher than in the realms of general education or general learned abilities.

Indicators: Inputs and Outputs

The question was raised principally in light of the paucity of such indicators in The Condition of Education, the annual report to Congress on the health of American education prepared by the National Center for Education Statistics. After all, the current concern for quality in American education requires data that can answer the questions, "What kinds of knowledge are students bringing to the workforce and to society?" and "What progress are we making in improving that knowledge?" Congress, State legislatures, and employers all have an intrinsic interest in answers to these questions.

But the data we customarily collect principally concern the "inputs" of higher education (percentages of faculty with terminal degrees, average SAT scores of entering college freshmen, numbers of degrees and courses offered, numbers of volumes and journals in university libraries, etc.). Data on the academic "outcomes" of higher education, such as those found in The Condition of Education or the annual Digest of Education Statistics are limited to mechanical measures (e.g., numbers of degrees awarded, persistence rates, changes in educational aspirations, trends in applications to graduate and professional schools), none of which indicate what students actually learn in college. Data extracted from opinion surveys concerning the quality of students entering graduate or professional programs, while accessible (e.g., Anderson, 1984; Atelsek, 1984), also do not tell us what college students learn.

What we tried between 1985 and 1988 in The Condition of Education was a group of indicators based on changes in scores on the Graduate Record Examinations, both General and Subject Tests. As the staff member responsible for preparation of this indicator, I chose the Standard Deviation Unit (SDU) as the metric, and, in the most recent versions of this presentation (NCES, 1988), distinguished between long-term (since 1964) and intermediate-term (since 1976) trends. The SDU is a far more accurate metric than mean score to measure change in the performance of a large and varied population over a period of time, and the Graduate Record Examinations (unlike the College Board Achievement Tests given to high school students) have never been rescaled, hence can be used with some historical confidence as the grounds of time-series data necessary to construct indicators (Adelman, 1985).

But there are three major limitations of indicators based on results of examinations such as the GREs, particularly the Subject Area tests, that lead us to this project: (1) non-generalizable content; (2) self-selected samples; and (3) inaccessible interpretation.

1. Non-Generalizable Content. Despite the fact that each GRE Subject Area test is governed by a Board of Examiners, all of whom are college and university professors appointed with the advice of their professional/learned societies, and despite the fact that this representative Board sets the technical specifications for the examination, the content of the test is not generalizable, that is, it may not reflect the actual undergraduate curriculum in that field as practiced at many institutions. In recent years, "content representativeness studies" of at least six GRE Subject Area tests have been conducted, demonstrating that, in the eyes of teaching faculty, the curriculum as practiced in their departments is not adequately reflected in the distribution of GRE test items (Oltman, 1982; Devore and McPeek, 1985). The content of the exams, then, more likely represents a "core" conception of a field, around which there may be many variations. These variations may be determined by the type, size, or even geographical location of an institution (Haswell and Lindquist, 1965), size and resources of departments, and different areas of faculty expertise and preferred approaches within departments. In this sense, the measurement is not flexible (Brown, 1970).

2. Self-Selected Samples. The sample of students taking the GRE Subject Area tests is driven, in part, by the changing admissions requirements of graduate programs and fellowship sponsors in the field, but is otherwise self-selected. Based on studies of those who took the GRE General Examinations (Verbal, Quantitative and Analytic), one can also infer that this group of test-takers is of higher ability relative to the general population of college graduates (Grandy, 1984). There may be a small portion of test-takers in some fields who sit for the exams as part of departmental program evaluations, but that is impossible to determine. In some fields, too, the number of GRE Subject Area test-takers is too small for meaningful interpretation of results.

3. Inaccessible Interpretation. Indicators should render complex phenomena accessible to a broad spectrum of users of information (NCES, 1985). When the language of an indicator--number, formula, words--is essentially foreign to that spectrum of users, the indicator does not fulfill its intended purpose. And when indicators of similar phenomena are expressed on different scales, they illustrate the difficulty of meaningful aggregation (Bell, 1973). No two GRE Subject Area tests use the same scale, hence mean scores are difficult to compare from test to test. That was one reason for choosing the Standard Deviation Unit as the metric for the one indicator we currently use. But despite its psychometric virtues, the SDU remains conceptually inaccessible to the broad spectrum of educators, policymakers, and the public. Due to their repetitive use in the national media, the metric of mean scores on the SATs have taken on the characteristics of a transparent public indicator such as the Dow Jones Industrial Average. Few users of these indicators can tell you precisely how they are generated, but nearly all users think they know what these indicators mean. This is simply not the case for the SDU any more than it is in the financial markets for the "Investors' Intelligence Overbought/Oversold Oscillator."

Given these constrictions on the task of constructing indicators from unobtrusive data in our de facto national examination program in the disciplines, we decided to ask the disciplines themselves a creative-thinking question:

> "How would you develop, validate, and produce broadly usable indicators of summative college student learning, covering content, methods, and modes of thought in X field (biology, business administration, anthropology, philosophy, etc.)?"

In July of 1986, the Office of Research issued a Request for Proposals (RFP) that elaborated on this question, and that set forward a series of requirements and tasks for the studies we would sponsor over an 18-month period. Because the question was speculative, and specifically did "not require the contractors to develop and validate tests or any other assessments," we anticipated studies of modest scope, and hence planned to sponsor as many as six under the same RFP.

The Competition and Its Results

While I have described the process and its results in detail elsewhere (Adelman, 1988), it is important to note here that we were calling for model building, not the implementation of models. We asked that each contractor conduct an inquiry "that reflects on the knowledge paradigm(s) of the field and current college curricula in that field, reviews the strengths and weaknesses of available tests and assessment methods in that field, and yields a description of how one would construct measurement(s) of 'summative' undergraduate learning in the field that select, weight, and otherwise account for both the generalizable portion of the curriculum and the special emphases of individual departments." We also asked that each contractor examine methods of assessment in his or her discipline as practiced in other nations and alternatives to conventional testing models of assessment, and that each contractor involve in the project representatives of appropriate professional or learned societies.

While the question and the tasks were posed from the perspective of public policy obligations of the Office of Educational Research and Improvement, there is no doubt that the project was intended to stimulate a process of reflection on the objectives of undergraduate study in the disciplines and the possibility of demonstrating the attainment of those objectives in economical, publicly accessible, and convincing ways. And, as the reader of the chapters in this volume will observe, the results of the project should prove more beneficial to individual institutions than to national data reporting.

We received 20 proposals in response to the RFP, and in fields ranging from nursing to foreign languages to computer science. The five winners of the competition whose final reports are presented in this volume were all in the basic sciences and applied sciences. Why? Did they understand the question better than did those in other fields? Is there more of a congruence between the quantitative nature of indicators and quantitative fields that leads the latter to a natural affinity with the task? No, because social indicators are not like mathematical constructs that assume "perfect" distinctions among categories and "perfect" homogeneity within them. (Roberts, 1974) Indeed, in order to explain the inperfect conceptual framework that often lies behind them, social indicators often use formulas and words in addition to numbers. (Kruskal, 1978)

Given the fact that we were explicitly interested in non-traditional approaches to the language of indicators, I think the answer has less to do with knowledge paradigms than with the organization of academic work in the disciplines. To put the case simply, and as Anthony Becher has demonstrated, the sciences and applied sciences are more "urban" in their organization, hence their practitioners are better networked than those in other fields, and can respond quickly and incisively to proposal initatives (Becher, 1989).

Consensus and Paradigm

The conventional wisdom would hold, however, that scientific fields with strong knowledge paradigms and a high degree of consensus on expected student learning outcomes were in a better position to address the central question convincingly. The conventional wisdom concerning other major areas of knowledge is that the humanities are pre-paradigmatic and many of the social sciences are governed by multiple paradigms (Ritzer, 1982). In either case, the result is a low degree of consensus, at least about the ends of undergraduate education.

The case, however, is not so neat. Computer science and mechanical engineering, two fields represented in this volume and in which consensus is confounded by other variables, should illuminate this issue.

Following a distinction made by Pantin (1968), computer science is a restricted field of inquiry whereas mechanical engineering, like literary studies or history, is unrestricted. That is, computer science is restricted in terms of the range of phenomena with which it deals. The problems confronted by computer scientists do not require them to use other fields of inquiry. This feature is a source of strength in the computer science knowledge paradigm, for, as Pantin observed, restrictions on classes of phenomena and numbers of variables only increase the deductive power of hypotheses. It should not be surprising that restricted fields tend to be quantitative, culumulative and dominated by universal laws. Mechanical engineers, on the other hand, frequently follow problems and questions into other disciplines, and can do so because the range of phenomena with which they deal is not as circumscribed. It follows that the knowledge paradigms of unrestricted fields are weaker. But that does not mean that consensus is more difficult to achieve.

Indeed, Powers and Enright (1986) have shown a remarkable degree of consensus among faculty in a notably unrestricted field, English, concerning the comparative importance of 56 discrete reasoning skills, 15 types of reasoning errors, and 25 types of "critical incidents" of student behavior that influence judgments of performance. To be sure, the focus of this study was on graduate education, and the topic was confined to reasoning skills. The fact that professors of literature evidenced a high degree of consensus, however, suggests that the general form of inquiry reflected in the reports of this volume can be pursued profitably in fields whose practitioners tell stories as well as those whose practitioners solve problems. The point is that the degree of consensus and the nature of the knowledge paradigm in a discipline simply should not preclude creative thinking about the evidence, the signs and traces, of college student learning.

To understand the limits of paradigm analysis, it may be helpful for the reader of this volume to consider the difference between indicators of knowledge and indicators of learning. Of the process of paradigm change in science, Thomas Kuhn stressed that the reigning theory itself determines the problems admissible for scientific investigation, the standards for investigation of those problems, and the organization of scientific work. (Kuhn, 1962) While Kuhn pointed out that scientific revolutions are ultimately canonized by the textbooks and modes of education in the sciences, he did not extend that analysis. What is known, however, is not necessarily what is learned, and the indicators of the two are rather different. This postulate, I submit, applies in all basic academic disciplines.

Think, for a moment, of the symbols we use to identify the advancement of knowledge. Some refer directly to the reality, some indirectly. Among the direct indicators are the number of entities identified in a field (species, elements, original artifacts, original texts) and the degree of detail in the description of knowledge in a field (a fine example is the growth of linguistics following Chomsky's articulation of generative grammar). Among the indirect indicators are the number of scholarly articles published in a field and the number of distinct disciplinary professional associations (and their academic journals) and types of degrees awarded in a field.

These indicators tell us something of what is known, and all of them presuppose expertise of the knower. But they do not represent the reality of how well the "what" is known by those who are not experts, those who are novices or journeymen, those who undergo educational initiation in a field. In other words, they are not indicators of learning.

It might be argued that some of these indicators would represent learning if the social context were different, if, for example, we lived in an extremely hierarchical society in which all knowledge was considered sacred and its acquisition limited to an isolated order of untouchable monks. But even such monks would be required to undergo educational initiation, and their masters would be eliciting representations of learning by the various means we now call "assessments."

The fact is, however, that our social and economic context, along with the values of a democratic society, call for the mass diffusion of knowledge and a considerable public investment in the process of diffusion. For us not to know how well the "what" is known by the mass would be analogous to investing in a financial instrument without any subsequent knowledge of its absolute or relative performance.

The indicators that might be constructed from the models presented in this volume would be navigational charts of what is being learned on the broader sea of knowledge. All the authors of the chapters that follow would agree that there are inherent problems with these charts, no matter how carefully crafted and discriminating the technology that produces them. One problem, for example, lies in the distinction between immediate learning and retained learning. If we hold the latter to be a more important outcome of our efforts in higher education, then we will use different types of assessments to elicit the signs and traces of learning than those normally used in the college classroom. Another problem concerns the sources of learning. That is, few modes of assessment can tell us whether the signs of learning we observe are rooted in what the student studied in college or in what the student gained in other contexts.

For these reasons, none of the models advanced in this volume propose a single indicator or a single source of information about student learning.

The Contents of This Volume

This volume consists of versions of the five project final reports, each underscoring a distinct approach to the task. Because the authors of these reports met twice during the course of their work, and because we invited interested parties from other federal agencies, higher education associations, and learned societies to those meetings to ask questions and offer advice, there are both common points of reference and self-consciously divergent paths. Certain phrases and arguments will reappear in a number of reports. As an editor, I do not deem these reappearances to be redundant, rather necessary rehearsals through which the authors indicate to the reader how they progressed from their reinterpretation of the task to the model of indicator-construction they recommend.

Each report offers a unique model. But while presented in the context of a specific discipline, none of those models is discipline-specific. For example, the model in Physics is applicable to any field dominated by textbooks in its undergraduate presentation. The model in Chemistry is applicable to any field with a history of licensure, certification, and/or accreditation. The model in Biology is applicable to fields in which sub-fields proliferate and, in large institutions, form their own departments. The point is that these reports are written not only for faculty in physics, chemistry, biological sciences, engineering, and computer science, but also for faculty in economics, nursing, history, psychology, and business, for example. The sections of Grandy's essay on computer science that deal with the nature of indicators or mastery-testing, and those of Warren's essay on mechanical engineering that analyze faculty expectations and faculty course examinations are, in fact, generic. They are for everyone concerned with the quality of student learning and assessment in the disciplines.

In many ways, the essays in this volume are about assessment, and assessment in the baccalaureate major in particular. Under this aspect of modelling, the reader will find some very creative approaches, along with examples of assessment tasks that can be translated into the regimens of other disciplines. Virtually all of the essays in this volume recognize that the methods of measuring student achievement are not limited to timed "paper-and-pencil examinations," let alone standardized multiple choice tests. All of them take into account the requirements of criterion-referenced and norm-referenced information (though with a heavier emphasis on the former). And all of them demonstrate how appropriate assessment data in their field can be used, first and foremost, by students and faculty in individual departments. Their recommendations in this regard are both technically exacting and innovative, hence exemplary.

The position of the Office of Educational Research and Improvement of the U.S. Department of Education with respect to the models presented and the steps recommended for realizing these models should be underscored: the Departmental imprimature on this volume implies neither endorsement nor recommendation. OERI's responsibility is to sponsor research and provide information that may help improve American education. We can only commend this volume to its readers as worthy of serious consideration.

References

Adelman, C. (1985). The Standardized Test Scores of College Graduates, 1964-1982. Washington, D.C.: The National Institute of Education in cooperation with the American Association for Higher Education.

Adelman, C. (1988). "Monstrous and less erroneous pictures: indicators of learning in the disciplines." Liberal Education 74 (3), 17-22.

Anderson, C.J. (1984). Student Quality in the Humanities: Opinions of Senior Academic Officials. Washington, D.C.: American Council on Education.

Atelsek, F.J. (1984). Student Quality in the Sciences and Engineering: Opinions of Senior Academic Officials. Washington, D.C.: American Council on Education.

Becher, A. (1989). Academic Tribes and Territories. Philadelphia: Open University Press.

Bell, D. (1973). The Coming of Post-Industrial Society. New York: Basic Books.

Brown, D. (1970). "A scheme for measuring the output of higher education," in B. Lawrence, G. Weathersby, and V.W. Patterson (eds.), Outputs in Higher Education: Their Identification, Measurement and Evaluation. Boulder, Colo.: Western Interstate Commission for Higher Education, 27-40.

DeVore, R. and McPeek, M. (1985). Report of a Study of the Content of Three GRE Advanced Tests. GREB No. 78-4R. Princeton: Educational Testing Service.

Grandy, J. (1984). Profiles of Prospective Humanities Majors: 1975-1983. Princeton: Educational Testing Service (Final Report for NEH Grant #OP-20119-83).

Haswell, H.A. and Lindquist, C. B. (1965). Undergraduate Curriculum Patterns: A Survey of Baccalaureate Programs in Selected Fields. Washington, D.C.: U.S. Office of Education.

Kruskal, W. (1978). "Formulas, numbers, words: statistics in prose." The American Scholar 47 (2), 223-229.

Kuhn, Thomas S. (1962). The Structure of Scientific Revolutions. Chicago: University of Chicago Press.

National Center for Education Statistics (1985). Indicators of Education Status and Trends. Washington, D.C.: U.S. Department of Education.

National Center for Education Statistics (1989). The Condition of Education, 1988, Volume II, Postsecondary Education. Washington, D.C.: Author.

Oltman, P. K. (1982). Content Representativeness of the GRE Advanced Tests in Chemistry, Computer Science, and Education. GREB 81-12p. Princeton: Educational Testing Service.

Pantin, C.F.A. (1968). The Relations Between the Sciences. Cambridge (England): Cambridge University Press.

Powers, D. E. and Enright, M. K. (1986). Analytical Reasoning Skills Involved in Graduate Study: Perceptions of Faculty in Six Fields. GREB 83-23p. Princeton: Educational Testing Service.

Ritzer, G. (1982). Sociology: A Multiple Paradigm Science. Boston: Allyn & Bacon.

Roberts, M.J. (1974). "On the nature and condition of social science." Daedalus 103 (3), 47-64.

Models for Developing Computer-Based Indicators of College Student Learning in Computer Science

Jerilee Grandy
Educational Testing Service

The purpose of this project was to develop a model of one or more indicator(s) of summative undergraduate learning in the discipline of computer science.[1] The inspiration for the project grew out of the current concern for quality in American higher education and, in part, out of dissatisfaction with the type of data available to policymakers and academic leaders.

To assess and improve the quality of higher education, policymakers and academic leaders must have relevant, well-defined, valid, and reliable indicators of student learning. Most of the statistics available to them are based on data that exist for some other purpose. While those statistics may reflect important aspects of student learning, rarely, if ever, do they convey the specific information needed for policy or program improvement. Educational decisionmaking is too important to be based on vaguely related statistics that just happen to exist. It should be based on information gathered specifically for that purpose.

For this project, we focused on how we would develop appropriate indicators of student learning in the major field of computer science at a 4-year college or university. While any major field of study might have served equally well as a model, we chose computer science because it is a relatively new and popular discipline, its content may be changing rapidly with advances in technology, and from our reviews of college catalogs, it appears to be a field that has content that varies considerably from one institution to another. These characteristics would pose a challenge to a system producing indicators because that system would have to be sensitive enough to reflect differences in departmental emphases and flexible enough to be modified and revised easily.

---

[1] The number of friends and colleagues who have taken an interest in this project and donated their time and ideas testifies to its potential value in education. Without the help of outstanding computer scientists as well as experts on computer-interactive testing, this project would have been impossible.

I wish to thank the members of my advisory committee, Professors Rafael Alonso, Kenneth Supowit, and Brian Reiser from Princeton University, and Professor Louis Sternberg from Rutgers University. Special thanks go to Ken and Rafael for contributing sample test items to illustrate important skills that require computerized testing.

Among the many people who provided helpful reviews of the earlier sections of the manuscript was Dr. Emilie Roth of the Human Sciences Department of Westinghouse Research and Development Center in Pittsburgh.

Additional thanks go to my colleagues at ETS who helped me to put my wild ideas into perspective and to those who reviewed parts of the manuscript. They included Randy Bennett, C. Victor Bunderson, Drew Gitomer, Roger Kershaw, Juan Moran-Soto, Eldon Park, Kathy Sheehan, and Martha Stocking.

The Nature of an Indicator

An Epistemological Beginning

To the chemist, litmus paper is an indicator of acidity. To the meteorologist, a falling barometer and a stratocumulus cloud cover are indicators of an impending storm. To an engineer, a non-zero reading on an ammeter indicates that current is flowing through the circuit in which the ammeter is installed. To Isaac Newton, a falling apple was an indicator of the earth's gravitational field.

In the social and behavioral sciences, we also speak of indicators: social indicators, economic indicators, and education indicators. In these sciences, unfortunately, it is not always clear what is an indicator of what. From common usage of the word, an indicator must indicate something. That "something" may be a type of event observable in the future, or it may be a construct, that is an abstraction that is not observable in itself but consists of many elements, generally too numerous to list. Examples of constructs include intelligence, knowledge, ability, attitude, socioeconomic status, poverty, happiness, and health. Similarly, the indicator itself may be observable or it may be a construct. The distinction between observables and constructs is important to make, not only so that we are clear about the nature of the indicator and how to generate it, but because the distinction has important implications for establishing the validity of the indicator.

Test scores in mathematics (observables) are indicators of mathematical knowledge and skills (constructs). Education indicators are nearly always indicators of constructs, and those constructs are rarely well defined. The result is that the linkage between an indicator and what it allegedly indicates is often tenuous. Consider as an example the now infamous decline in Scholastic Aptitude Test (SAT) scores between the 1960's and the 1980's. SAT scores (observables) are often cited as indicators of the general health of secondary education (a vague construct, at best). The SAT is designed, however, as a predictor of success in college, not as an indicator of high school outcomes. But few people see the SAT merely as a predictor of college performance.

When SAT score averages declined, virtually everyone--educators, parents, and students themselves--became alarmed. The score decline allegedly "indicated" a corresponding decline in the quality or "general health" of secondary education. What did that mean? Neither educational quality nor general health are well-defined constructs. Aside from test scores, what was really declining? Was it factual knowledge of a specific type, reasoning ability, motivation to take the test, physical health, or something not related to the general health of eduation at all, such as a population change? Because the SAT is not a diagnostic test, SAT scores are not indicators that are clearly linked to well-defined components of student learning or student behavior. There was no obvious way to take corrective action.

Without knowing what SAT scores or the score decline indicated, educators nevertheless launched a major effort to identify the causes of the decline so they could correct it. Committees convened to look for "explanations." The public speculated. The media blamed lack of discipline in schools, teacher incompetence, parental indifference, drugs, the Vietnam War, atomic fallout, disintegration of the family, and the tests themselves. Because SAT scores are not designed to diagnose particular insufficiencies in student learning, searching for the causes of those unknown insufficiences was not easy. To look for the cause of a change in an indicator without knowing what the indicator indicates can be truly an exercise in futility.

If an indicator can be closely linked with specific elements of student learning, causal connections are more likely to be evident and remediation is more promising. Students in a logic class, for example, who consistently confuse "not all true" with "all false" can generally profit from a review of quantifiers. The confusion of the same two logical concepts could occur on a reading test, but it would be likely to go undiagnosed. A specific type of reasoning error would be reflected in the total reading score, but the information necessary to help students clarify their confusion would be missing from that score. The understanding of quantifiers is a more narrowly defined construct than is reading skill. Indicators of narrowly defined constructs are more useful to educators and decisionmakers than are indicators of broader, vaguely defined constructs.

Aside from the importance of linking an indicator to a well-defined construct, there are other conditions that an indicator must satisfy. An indicator is a variable that may be descriptive or quantitative. The race or gender of a student is descriptive. The number of years a student spends earning a bachelor's degree is quantitative.

An education indicator may refer to an individual or to an explicitly defined group such as a class, department, institution, State, or the Nation. At the individual level, indicators are used by teachers and administrators to make decisions regarding individual student admission, placement, advancement, remediation, granting of honors, and graduation. Receiving a science fair award in high school may be a descriptive indicator of whether a student will perform well in college science courses. A student's grade in calculus is one quantitative indicator of whether that student is well enough prepared to complete a course in differential equations.

At the group level, an indicator is a statistic. If it is a statistic based on a descriptive variable, the indicator will generally be expressed as a proportion or percentage of the group who lie within a category. An example of a descriptive statistical indicator is, "51 percent of our graduating class is female." If the indicator is a statistic based on a quantitative variable, it is likely to be expressed as a mean or median. An example of a quantitative statistical indicator is the average number of years students spend earning a bachelor's degree. That figure could be one economic indicator used in assessing the monetary value of a college education.

In summary, the types of indicators generally used in education are shown diagramatically in the chart that follows. The two broad categories of indicators are descriptive and quantitative. Each type may be applied to an individual or a group. When applied to a group, an indicator is a statistic, and the nature of that statistic depends upon whether the indicator is a categorical or quantitative variable. The examples include some of the most commonly used indicators.

There are additional conditions that an indicator must satisfy. An education indicator can be misleading or meaningless unless the nature of the reference group is clearly defined. A major problem in explaining the SAT score decline is that the population taking the examination is undefined, and there is no reason to believe that the population is comparable from year to year. Changes in the characteristics of the population choosing to take the test could account for the decline. If so, the decline may not have been an education indicator at all, but a reflection of new reasons for taking the test. (For a thorough review of the literature on the SAT score decline, see Waters, 1981).

Types of Indicators Useful in Education

	Descriptive		Quantitative	
Applies to:	Individual	Group	Individual	Group
Representation:	Categorical	Statistical (percent of group in category)	Scaled	Statistical (Mean or Median)
Examples:	Race Gender Pass/fail	percent Black percent Female percent Passing	Income Grade Test score	Median income Mean grade Mean score

The discussion of indicators thus far has focussed on types of indicators and the characteristics required for them to be technically sound. But there are other characteristics that indicators should possess if they are to be practical to collect and useful to educational leaders.

An education indicator should be readily understood by the leaders and policymakers who must use them. An indicator that is generated by a series of 19 stepwise regressions then standardized with a mean of 30 is not likely to be meaningful to anyone but the statistician who produced it. An indicator that cannot be explained in a simple sentence may be misused or used incorrectly.

To be practical, an education indicator should be easily generated. It should be based on data that are easily collected. Indicators that require excessive time and expense to collect will be available only from sources that are sufficiently motivated to participate in their collection. The result will be incomplete data providing indicators based on a population that is difficult or impossible to define.

Indicators should be easily modified so that they reflect currently relevant aspects of education. A test that is quickly outdated and requires a million dollars for its revision is not practical. Ideally, methods for updating an indicator should be specified in the design of the indicator system.

While these characteristics of education indicators are not exhaustive, they should help to set the stage for the design of specific types of education indicators, namely, indicators of student learning. (For additional discussions of education indicators, see Oakes, 1986, and Stern, 1986.)

Indicators of Student Learning

"Student learning" is a construct, whether we speak of student learning in general, student learning of computer science, or student learning of processors and control units. The more narrowly defined the construct the more specific and useful we will find the indicator information to be.

For the purposes of this study, we define undergraduate student learning to be the "increase in the knowledge, skills, and abilities of a student or group of students between the time they enter college and the time they graduate." Undergraduate student learning may be restricted to a particular discipline, such as the major field, or it may refer to a limited type of learning, such as cognitive learning. Without these specified restrictions, it includes both cognitive and noncognitive learning.

Summative learning, in this report, is a broad construct referring to the sum total of all learning. "Summative undergraduate student learning in computer science" is the total of all student learning in computer science between college admission and graduation.

Indicators of student learning in computer science must satisfy the requirements for acceptable indicators as described in the previous section. In addition, the realm of computer science as a discipline must be defined in terms of its boundaries and components. An indicator of summative learning in computer science should be based on a definition of summative learning that specifies the boundaries of computer science and weights its components in an acceptable manner.

Because indicators must represent student learning over a specified period of time, measures must be taken at least twice: upon admission and upon graduation. Any indicator of student learning will involve computing a difference between those two measures. Put somewhat differently, an indicator of student knowledge, skills, and abilities must be generated early in the student's freshman year, or upon admission (t_1), and again around graduation (t_2). The difference between the value at(t_1)and the value at(t_2)is an indicator of student learning, though that difference may be more complex than an arithmetic difference between two test scores. As seniors, for example, students may use a more efficient or elegant algorithm to solve a problem than they used to solve the same problem as freshmen.

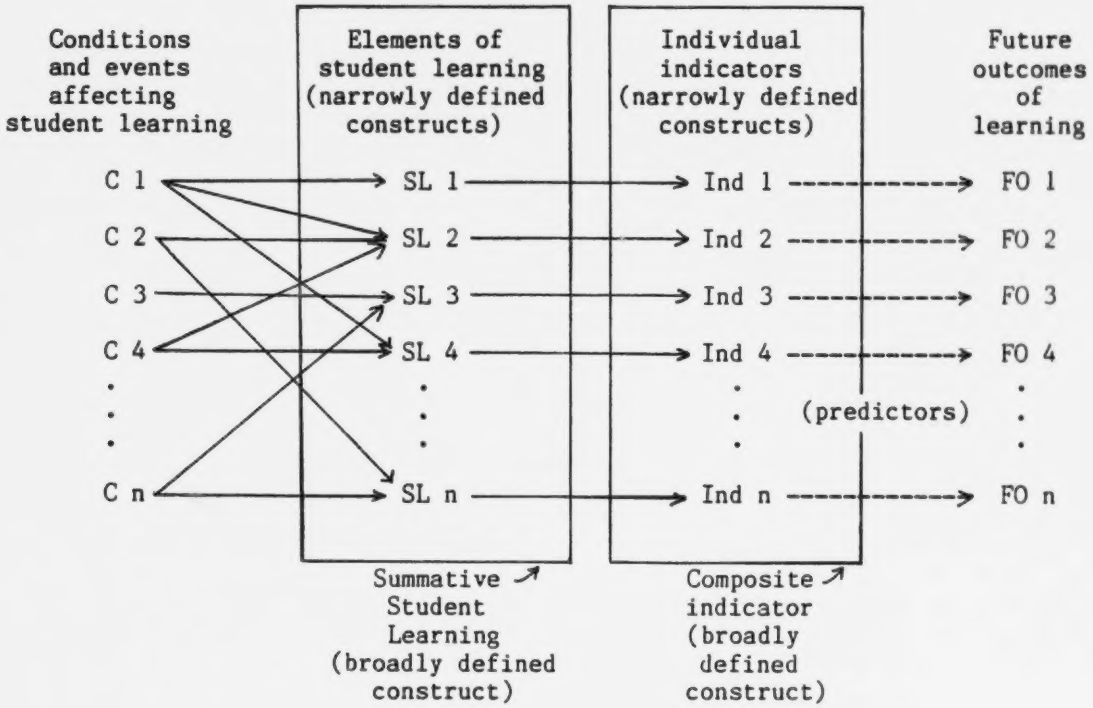
The following chart illustrates the relationships between indicators and student learning. It also sets forth model linkages between constructs and events affecting both student learning and future outcomes of student learning.

Computer-Interactive Testing

Any method of testing in which the examinee interacts directly with the computer is called computer-interactive testing (CIT). Information elicited by the computer's testing program reflects the examinee's knowledge, skills, and abilities at the time of testing. By testing the student periodically, we can observe or compute increases in knowledge, improvement in skills, and the development of new abilities. This growth constitutes student learning, and can be

Relationships between Indicators and Student Learning

(solid arrows represent apparent causal connections;
broken arrows represent predictions)


Conditions and events affecting student learning
Elements of student learning (narrowly defined constructs)
Individual indicators (narrowly defined constructs)
Future outcomes of learning
C 1
C 2
C 3
C 4
C n
SL 1
SL 2
SL 3
SL 4
SL n
Ind 1
Ind 2
Ind 3
Ind 4
Ind n
(predictors)
FO 1
FO 2
FO 3
FO 4
FO n
Summative Student Learning (broadly defined construct)
Composite indicator (broadly defined construct)

expressed quantitatively or descriptively. This expression of student learning is thus an indicator of student learning. Later in this report we will use the term computer-based indicator to refer to an indicator generated by or derived from CIT.

CIT has a number of advantages over paper-and-pencil testing:

1. The format of test questions is not limited to multiple choice. Answers to computational problems can be typed in. Diagrams can be drawn or altered, circuits can be traced, and components in circuits can be moved or replaced. To test programming skills, the examinee may be asked to write a program segment or to analyze or modify one presented by the testing program. This capability is one of the greatest strengths of CIT and provides the most creative possibilities.

2. The testing program offers a range of immediate scoring formats for the examinee. These formats may include a completely individualized report with diagnostic information, explanation of errors, and remedial instruction. They may present normative results that compare the examinee with others who took the test in previous years throughout the country, within his or her college, or both. If the test is "criterion-referenced" and a cut-score has been set, one can draw upon a format that indicates whether the examinee has passed or failed. (These distinctions are discussed in most textbooks on measurement. See, for example, Anastasi, 1988.)

3. The test battery can be regularly updated. Faulty or obsolete items are easily replaced. Entire content areas, or modules, can be added or deleted. Thus, if a test of skill in a specific programming language is desired ten years from now, that test can simply be added to the battery. The testing program itself can be an evolving system as methods of assessment become more sophisticated, computer technology changes, and emphases on different content areas change.

4. Test security problems are mitigated. The computer can choose questions for a particular examinee at random, thus making it highly unlikely that two examinees will be asked the same set of questions. Only the test disks themselves and the algorithm for choosing questions must be kept secure.

5. Computerized tests are easy to administer, particularly to computer science students who have ready access to hardware and enough experience in its use that they are unlikely to experience "computer anxiety." A test administrator requires virtually no training because the program itself guides the examinee.

6. No special computer hardware is required. A readily available microcomputer, with a printer, graphics monitor, and graphics board will serve most testing purposes.

7. Studies by ETS and the College Board indicate that students prefer computerized tests (Ward, Kline, and Flaugher, 1985).

CIT is not limited to any particular item format, though the psychometric properties of some item types may not be known and may have to be developed. The multiple-choice format has been studied the most, and both true-score theory and item response theory (IRT) have been highly developed to deal with the psychometrics of multiple-choice tests. Free-response items (also known as constructed-response items) that can be scored by computer are also being actively researched and applied. In most instances, however, constructed response items are being used to provide descriptive diagnoses of examinee errors rather than a quantitative score.

Computerized Adaptive Testing

To assess student learning with adequate reliability in all of the areas of computer science, a paper-and-pencil test would have to be much longer and require far more of the examinee's time than is realistic. (For a discussion of the relationship of test length and reliability see a standard textbook on measurement such as Anastasi, 1988). If individual students' scores must be reliable, or if score averages for small departments must be reliable, a test must be longer than it would have to be if only score averages for very large groups were required. Similarly, if reliable subscores for specialized content areas are required, each of the content areas must be longer than it would have to be if it were only contributing to a total score.

Computerized Adaptive Testing can avoid the psychometric problems just mentioned by replacing the paper-and-pencil test with a computer-interactive test. Adaptive testing applies the statistical methods of IRT to tailor the difficulty of a test to the skills of individual examinees (See Wainer, 1983, and Lord, 1980). As a result, it can provide very efficient measurement across a broader range of levels of skills than can be achieved with conventional paper-and-pencil tests. Using the computer enables a rapid determination of which questions should be given to an individual; it also permits immediate reporting of the test results on a scale identical to the kinds of scales to which examinees are accustomed.

The simplest form of an adaptive test works in the following way:

Step 1. The computer chooses at random one of the middle- difficulty-level questions and presents it to the examinee.

Step 2. Depending on whether or not the student answers correctly, the computer randomly selects the next item either from the easiest or the most difficult questions.

Step 3. The computer continues to monitor responses and chooses from questions of appropriate difficulty until it "zeroes in" on the student's skill level.

This description of an adaptive test is an oversimplification of the process. In actuality, the first item may be on a difficulty level predetermined by some prior indicator of the examinee's ability. A freshman computer science major, for example, might be presented with an easy item to start, while a graduating senior may be presented with an item of above average difficulty. Another variation in the three-step process is that instead of branching to an easier or more difficult item based on the examinee's response to a single item, the program may present several items of the same difficulty before deciding whether to branch to a different level of difficulty. The basic logic is the same, however. What is unique to an adaptive test is that the items selected for presentation to the examinee are adapted to the demonstrated knowledge and skill level of the examinee.

An adaptive test can have a great many advantages over the traditional paper-and-pencil test:

- It can provide fine discrimination over a wide range of ability levels. In constrast to a traditional test which has high precision near the average test score, an adaptive test can be designed to have the same precision for examinees at all ability levels.

- Testing time can be considerably shorter than with a paper and pencil test. The program eliminates questions that fall outside the examinee's ability range, therefore determining the skill level precisely with a minimum number of questions. The examinee's score is determined not by how many questions are answered correctly but by which questions are answered correctly. Gialluca and Weiss (1979) and Maurelli and Weiss (1981) found that they could reduce the total length of a biology test from 16 percent to 30 percent with virtually no loss in psychometric information.

- An adaptive test can be adapted not only to the examinee but to the institution as well. One institution may decide to have its students take some subtests and to bypass others, or it may instruct the examinees to skip a subtest in an area in which they have no knowledge.

- With an adaptive test, students are generally not bored by having to answer questions that are too easy, and they are not anxious or discouraged by attempting large numbers of items that are too difficult.

Some adaptive tests are currently in regular use. The College Board's Computerized Placement Tests in Reading Comprehension, Sentence Skills, Arithmetic, and Algebra are examples. These untimed computerized tests automatically produce scores and a variety of score reports and summaries. They can discriminate better and more reliably than paper-and-pencil tests but take only about half as much time to administer because they contain only 12 to 17 items each (College Entrance Examination Board, 1985).

There is a computerized version of the Armed Forces Vocational Aptitude Battery (ASVAB). The Psychological Corporation has developed an adaptive version of the Differential Aptitude Test (DAT) for administration on Apple II computers (Psychological Corporation, 1986). ETS is developing other adaptive tests, especially for use in licensing and certification.

The use of adaptive testing is not without problems, however, and the design of adaptive tests requires special skills and demands great care. Wainer and Kiely (1987) have discussed three technical problems: context effects, lack of robustness, and the order of items by difficulty.

1. Context effects refer to any effects on item performance caused by an item's relationship to other items in the test. If all examinees are presented with the same items, the context effects are assumed to be the same for everyone. But with different students receiving different items, the context will not affect everyone similarly. One example of this phenomenon arises when the information required to answer an item is contained within an earlier item. If some students have not been presented with the same earlier item, the successive item will be more difficult for those students.

Similarly, an item may be more or less difficult depending on its order of presentation. A number of studies have found differences in difficulty parameter estimates for items as a function of their location in the test (Whitely and Dawes, 1976; Yen, 1980; Eignor and Cook, 1983; Kingston and Dorans, 1984).

A third type of context effect arises if a particular theme or subject appears repeatedly. In a traditional paper-and-pencil test, test developers now avoid unbalanced content, especially on socially sensitive subjects. For example, if a reading passage uses a male first name for a person in a technical job, an item writer will take care to use a female name in another item having a person with a similar role. Likewise, if one sentence completion item refers to a traditionally female type of recreation, such as dance, another item may refer to a traditionally male sport, such as basketball.

In adaptive testing, because not all examinees are presented with the same items, one person may by chance get all mostly "male" items, while another gets mostly "female" items. A similar kind of imbalance would occur if one examinee got two successive items where the main character had an Hispanic name, or several vocabulary items drawn from literature and none from science.

2. Lack of robustness essentially means that if an item is flawed in some way, its detrimental impact is greater on an adaptive test than on a traditional paper-and-pencil test of greater length. This is because the shorter test lacks the redundancy inherent in a longer, conventional test. If an item fails to perform as expected, its detrimental impact on validity may be considerable.

3. A number of studies have indicated that the ordering of items by difficulty has an effect on student performance (Mollenkopf, 1950; MacNicol, 1956; Sax and Carr, 1962; Hambleton, 1968; Monk and Stallings, 1970; Towle and Merrill, 1975). These studies have shown that when items are arranged in order of decreasing difficulty, instead of the typical order of increasing difficulty, the overall effect is to increase the difficulty of the test, probably by increasing anxiety and frustration. In adaptive testing, the lower ability examinees are first presented with an item of medium difficulty, which for them is of high difficulty. The computer then presents them with successively easier items. The net effect for those examinees may be a test containing items of decreasing difficulty.

When a paper-and-pencil test is modified for use as a computerized adaptive test, its validity must be re-established. In a factor analytic study, Green (1986) showed that when some paragraph comprehension items were modified to facilitate computer presentation, they came to look more like word knowledge items. Thus by adapting paper-and-pencil items to ones that can be presented by computer, we risk loosing what we originally intended to measure.

Psychometricians have other concerns about adaptive testing. The use of change scores to measure achievement over time assumes that the pretest and posttest are measuring the same skills or knowledge, that is that the factor underlying changes in performance is invariant over time. Gialluca and Weiss (1981), however, found that this was not always true. They found that the factor structure of achievement in a biology course was not the same before instruction as it was several weeks after instruction, whereas in a mathematics course, the factor structure remained the same over a 10-week period.

While there are still problems in the application of adaptive testing, the problems do not preclude its further development. Wainer and Kiely (1987) and their colleagues have been applying a multistage fixed branching adaptive-test model that substitutes multi-item "testlets", or very short tests, for single items. They define a testlet as a "group of items related to a single content area that is developed as a unit and contains a fixed number of predetermined paths that an examinee may follow." Because each item is embedded in a predeveloped testlet, it essentially carries its own context with it.

The path that an examinee follows through the testlet may be a hierarchical branching scheme to items of greater or lesser difficulty depending on the previous responses. Each path leads to one of a series of ordered score categories. Alternatively, the examinee may work through a linear path of a number of items that are presented to all examinees. The testlets themselves can then be linked either hierarchically or linearly, depending on their intended use.

According to Wainer and Kiely (1987), the fixed branching testlet avoids, or at least minimizes, the problems inherent in the variable branching adaptive test models currently under development. It provides only a limited number of paths for examinees to follow, and therefore, if the test developer constructs each path carefully, he can avoid most of the problems identified earlier.

While many researchers and test developers are confident that testlets are psychometrically sound, some are concerned that an occasional problematical item will inadvertantly be used, and examinees' scores will be affected. Some test developers are also skeptical about public acceptance. Their specific concern is that in spite of psychometric arguments (which are understood only by psychometricians), the public will reject on grounds of intuition and common sense the notion that a test (or a test section) having only 8 or 10 items can possibly measure their knowledge of a subject. Furthermore, if one examinee takes different test items than another examinee, it will be difficult to explain in terms that are understandable to the public just how the test items are calibrated in order to make the two examinee's scores comparable.

Many people know that short tests, in general, are not very reliable. Convincing the public that some short tests are okay and others are not may be a difficult task. This problem is of particular concern to members of the legal profession, such as Randolph Reaves, leading authority on the legal aspects of certification and licensing and author of the most authoritative publication in the field, The Law of Professional Licensing and Certification. The legal concern is that an examinee who fails a test may argue in court that the test was too short to assess adequately what he or she knows. A judge, who may well listen to the statistician's explanation of item response theory, will still be likely to form his or her judgment on what appears reasonable and on the precedent of requiring that tests be long in order to be valid and reliable.

While there are still problems in adaptive testing, including the use of testlets, many relevant advances are being made in statistics and psychometrics. Research on new IRT scoring models and on extensions of Bayesian statistics are beginning to obviate the problems discussed above.

Computerized Mastery Testing

Mastery testing is a form of criterion-referenced testing. In criterion-referenced testing, a cutting score is set to indicate whether the examinee passes or fails. It may be set low to define minimal competency, or it may be set high to define mastery. In either case, it produces only a two-valued score rather than a continuous scaled score. When the cutting score is set to define mastery, it is called a mastery test.

Conventional paper-and-pencil mastery tests apply a cutting score above which the examinee passes and below which the examinee fails. An examinee who scores very near the cutting score has a higher probability of being misclassified due to measurement error than does an examinee whose score is very high or very low.

With very recently developed Computerized Mastery Testing (CMT), the computer presents testlets to the examinee sequentially. After the examinee completes each testlet, the program estimates whether the examinee is a master, non-master, or needs further testing, and it computes the loss associated with misclassification. Examinees who are very near the cutting score are given additional testlets until the probability of misclassification is sufficiently low.

Sheehan and Lewis (ETS, 1988) developed the Bayesian statistics needed for this CMT decision model and then combined those statistics with IRT scoring to design the Prototype Seismic Knowledge Computerized Mastery Test. This test is one division of the Architect Registration Examination which ETS developed for the National Council of Architectural Registation Boards (NCARB). Initial analyses suggest that by using the CMT rather than a paper-and-pencil test, the standard error of measurement at the cutting score can be reduced by about 18 percent. Not only are errors of misclassification reduced, but for examinees who score some distance from the cutting score, the test is considerably shorter than it would normally have to be. Some examinees take as few as 20 items. Those nearer the cutting score take a maximum of 60 items. The Seismic test also illustrates a new application of testlets that avoids the problems of adaptive testing mentioned earlier.

Constructed-response tests

One of the greatest advantages of CIT is that it does not limit testing to multiple-choice format. While paper-and-pencil tests also permit constructed-response formats, scoring generally requires human readers, and the task of reading examinations is time consuming and costly. Some of the most advanced research into CIT, therefore, is employing expert systems, or knowledge-based systems, to score student responses and diagnose errors. Much of this work is growing out of current research on Intelligent Tutoring Systems.

As part of a large Air Force-sponsored effort to understand, assess, and train electronics troubleshooting skills of technical personnel, Gitomer and his colleagues at ETS have developed a Logic Gate Tutor (Gitomer, 1987). The basis for assessment within this system is a form of latent class analysis: a mathematical tool that is useful for assigning individuals to qualitatively different groups based on the patterns of their responses. The diagnostic information derived from each assessment module is used to direct the instructional sequence. Individuals receive instruction only in those areas for which they have exhibited weakness.

One of the strengths of this kind of testing is that it identifies errors that are attributable to qualitative (i.e., descriptive) deficiencies in the learner's knowledge base rather than assuming a unidimensional model of measurement. Instead of reporting how much a student knows about a subject as a whole, it can report in which areas the student shows mastery and in which areas the student is deficient.

While the purpose of this and other similar tutoring systems is instruction, a system with diagnostic assessment components alone could provide useful indicators of student knowledge at any step on the learning ladder.

Anderson and his colleagues at Carnegie Mellon University have successfully designed a tutor, based on a learning theory known as ACT*, that teaches LISP (a programming language used to design expert systems). The tutor provides instruction in the context of problem solving, has the student generate as much of each solution as possible, provides immediate feedback on errors, and finally, represents the structure of the problem for the student. Anderson and Reiser (1985) showed that students using the tutor performed better than students in a standard classroom and nearly as well as students with personal tutors.

The key to the tutoring program's success is its ability to fit each student response into a model of correct and incorrect methods for solving a problem. The tutor must be able to analyze each portion of the student's solution in order to diagnose errors and to provide guidance. This process of understanding the student's behavior as it is generated is called "model-tracing" (Reiser, Anderson, and Farrell, 1985). By following a student's path through a problem, the tutor always has a model of the student's intentions. According to Johnson and Soloway (1984), inferring intentions is necessary for responding appropriately to students' misconceptions about problem solving.

Gong and Sleeman (1985) contend that most systematic errors in programming are due to misconceptions held by programmers who have incomplete or incorrect knowledge. Studies of novice programmers at the high school level have shown that a large number of programming errors may be attributed to misconceptions of what computers can do or to imperfect understanding of the syntax or semantics of the programming language. This was found to be true both in Basic (Putnam et al., 1984) and in Pascal (Sleeman et al., in press).

Johnson, et al (1983) have described a scheme for categorizing bugs in novice programs by identifying the misconceptions underlying the bugs. Following their work, Spohrer et al. (1985) have identified more programming bugs and classified them in catalogs that can be used by other researchers. Along with developing "bug catalogs," researchers at Yale have designed and implemented a program called PROUST which identifies non-syntactic bugs in programs written by novice Pascal programmers (Johnson, 1985). MicroPROUST is a prototype microcomputer version of PROUST. MicroPROUST has a knowledge base containing information about two easy problems on the College Board Advanced Placement (AP) Computer Science Test. The information is in a coded version of the English text of each problem. MicroPROUST's knowledge base divides the problems into subcomponents and analyzes solutions one subcomponent at a time, looking for templates in its knowledge base against which it can match portions of the examinee's solution.

A recently completed validity study comparing the judgments of MicroPROUST with the judgments of human readers on the two AP items has shown MicroPROUST to perform impressively on the solutions it could analyze but unable to grade a significant portion of the solutions (Bennett, et al, in preparation). Nevertheless, the data seem to suggest that, given certain constraints, the system does as well as readers. While there are still some implementation problems, researchers working with MicroPROUST see it as proof that within certain limitations, a computer is interchangeable with humans in grading complex constructed responses.

Advances in CIT over the past few years have resulted in improved expert systems to analyze and "understand" students' solutions to problems, new formats for presenting test items using adaptive testing and testlets, and the development of new statistics and psychometrics to handle these innovations. CIT is no longer a thought experiment or a desktop computer game looking for players. Standardized tests and intelligent tutors are being marketed and used by the education community. From our analysis of the characteristics of indicators and the conditions they must satisfy, CIT appears to offer a new and better way to produce meaningful education indicators.

The Content Domain of Computer Science

The Need for Definition

A prerequisite to the design of a system to produce indicators of student learning in any discipline is the need to define the content domain of that discipline. In choosing computer science as the discipline to model, we were fortunate to have selected a field in which a considerable amount of curriculum analysis and test development had already been undertaken.

In many respects, computer science has just reached adolescence as a discipline. Descending from the families of electrical engineering, mathematics, and business, computer science has been struggling to gain its own identity and reputability while its older and better established relatives are charging it with inadequacy, inconsistency, and lack of depth.

Some members of the IEEE Computer Society and the Association for Computing Machinery (ACM) report that they encounter computer science graduates unable to handle the professional responsibilities implied by their degree (Mulder and Dalphin, 1984). Their perceived deficiencies, however, may be attributable not only to true educational deficiencies, but to discrepancies between a particular college's interpretation of computer science and the standards and expectations of an employer or graduate school.

Depending upon its history, the curriculum for a degree program in computer science may differ greatly in its emphasis from one college to another. In a university with a large engineering department, computer science is likely to have grown out of electrical engineering, and it may still have a strong engineering component. The curriculum may contain required or optional courses in switching, computer network architecture, and digital computer design--all taught in the electrical engineering department.

In another college, possibly one emphasizing liberal arts, computer science is likely to have started as an area of specialization within the mathematics department and may still require linear algebra, discrete probability, combinatorial analysis, numerical methods, algebra, and more advanced applied mathematics courses.

A third type of computer science program is one with a strong business component. The curriculum requires courses in operations management, information systems, and systems analysis. COBOL may be a required programming language.

Quite clearly, an employer in a social science research environment who hires an A-student from the third type of institution will wonder why he or she, as a computer scientist, cannot write a simple FORTRAN program, or even use a canned program, to model some trend data. Likewise, the manager of a large business may regard a computer science graduate as incompetent if he or she has no knowledge of inventory control (but has developed an ingenious program for generating prime numbers).

In an attempt to set some standard guidelines, the ACM issued recommendations in 1968, and again in 1978, for the undergraduate curriculum in computer science (Association for Computing Machinery, 1983). Shortly after ACM published the 1978 revision, Ralston and Shaw (1980) charged that the new curriculum equated computer science with programming and that it lacked the mathematical content necessary for any discipline to call itself a science. They argued that mathematical reasoning plays an essential role in all areas of computer science and should be included in the curriculum from the beginning of the student's computer science education.

In another attempt to define the computer science curriculum as well as to set standards and evaluate computer science programs, the IEEE Computer Society and the ACM formed a Joint Task Force on Computer Science Program Accreditation. They too were faced with the task of defining the realm of computer science. In their view, computer science is one part of the whole field of computing sciences, which includes information systems, system analysis, data processing, and "computer science and engineering," in addition to pure computer science (Mulder and Dalphin, 1984). These areas overlap, of course, and the Joint Task Force defines computer science in such a way as to include some engineering, but to exclude all business applications, which include information systems, system analysis, and data processing. Not surprisingly, this definition of the realm of student learning in computer science has been criticized for being biased towards engineering, as well as being inflexible (Gibbs and Tucker, 1984).

The existence of these--and other--curriculum recommendations simply highlights the fact that experts cannot agree on the topics that constitute computer science and that there is a considerable difference in curriculum content across institutions. Despite such differences, all programs do have some common core knowledge, and it is this common core of learning that most achievement tests try to measure. For our purposes, however, it is insufficient to produce indicators of student learning only in the the common core areas of computer science.

In this project, we saw our task as defining the domain of computer science by placing all possible computer science topics into one of three categories: (1) topics that all experts agree lie within computer science; (2) elective topics and topics that are included in some curricula but not others; and (3) topics that may be related to computer science but are not actually part of any computer science curriculum. Category 1 comprises the common core of student learning in computer science, and category 2 contains the specialty areas that will provide indicators of the unique strengths of some departments. Topics in category 3 will be eliminated from a system of indicators of learning in computer science, but may, at some future time, be included in a system of indicators for a related discipline.

Defining Topics in Computer Science

We convened a four-member advisory committee to review existing test specifications and curriculum recommendations for the purpose of developing a satisfactory outline of the common core and speciality areas of undergraduate learning in computer science. Appendix A lists the names and affiliations of those members.

Most of our efforts in this phase of the project focussed on reviewing relevant test specifications and the ACM curriculum recommendations for topics that students were likely to cover in their undergraduate coursework. Not all existing test specifications in computer science were appropriate.

We found the specifications for the GRE Advanced Test in Computer Science (appendix B) to be highly relevant. This is not surprising, considering that the GRE Computer Science Test is designed to measure the student's mastery of subject matter emphasized in an undergraduate computer science program (Graduate Record Examinations Board, 1986).

Two of our committee members reviewed the specifications and suggested very slight modifications. The topics they would add, and the sections in which they would add them, are as follows:

IB: Loop invariants, invariant relations of a data structure
IC: Compiler techniques, code generation Parsers for formal languages
IE: Database management (e.g., transactions, serializability), Distributed systems
IIB: Computer arithmetic
IIE: Modeling and simulation
IIF: (new section): Parallel architectures
    Single-instruction-multiple-data
    Multiple-instruction-multiple-data
    Pipelined vector machines
IVA: Order notation
IVB: Numerical methods
    Analysis of roundoff error
    Numerical integration
    Matrix operations (e.g., Gaussian elimination)
    Interpolation and approximation
V: Special topics
VC: VLSI Design

The committee would also eliminate "Processors for formal languages" (III,A,5), "Correctness of programs" (III,B), and "Upper and lower bounds on the complexity of specific problems" (III,C,2). They would eliminate "abstract algebra" (IV,A,1), and they would also eliminate modeling and simulation, information retrieval, and data communications as special topics.

The ACM Recommendations for the Undergraduate Program in Computer Science were far more detailed than the GRE specifications. They included more topics, and classified them as core courses or electives, a distinction that was most important to our task of defining summative learning and special topics.

Because the ACM Recommendations and the Carnegie Mellon Curriculum (Shaw, 1985) were so thorough, the committee used them as a foundation for developing the final outline of topics. The decision as to whether a topic should lie within the common core of computer science or whether it should be a specialty area was not difficult to resolve. The mere fact of disagreement indicated that the topic would be included at some institutions and not at others. Since departmental emphases and differences were the reason for placing a topic in a specialty area, disagreement alone became the criterion for placing it there. Common core topics were ones that the committee all agreed should remain there.

The final topic outline (appendix C) divides the common core into four general areas: Programming, Software, Hardware, and Data Structures and File Processing. The special topics, that is, those that would distinguish one institution from another in its departmental emphases, are listed under Special Topics. There are nine special topics, including analysis of algorithms, languages, mathematics for computer science, computer engineering (hardware), software engineering, graphics, artificial intelligence and robotics, related management and information science, and related public policy.

College Catalog Search

As a final check on the completeness of the Computer Science Topics generated by a synthesis of other lists, we searched a small sample of college catalogs to see if any course topics were listed that did not arise in the test specifications or ACM curriculum. We conducted this exercise primarily to demonstrate a method for supplementing the topics we already had.

We selected, at random, a sample of twelve colleges offering "computer science" degrees. As a result of our decision to select only institutions that identified their degree as "computer science" (as opposed to "computer science and information systems"), the institutions included were all fairly large universities. If an actual assessment system were developed, we would use a more diverse sample.

Most of the institutions offered the same core courses, but at different levels, probably depending upon the academic preparation of the average freshman. The requirements for the major were different depending on the relative autonomy of the program vis-a-vis the mathematics or the engineering departments.

The twelve colleges we examined appeared to offer few unique computer science courses. Generally the differences among the programs lay in their relative emphasis on mathematics and in the acceptance of business courses. One university listed human factors, data security, and pattern recognition. If these topics were offered by the other colleges, it was not evident from their course descriptions.

Why can't we simply refer to catalogs to determine departmental emphases? Why is it necessary to test students directly? The main reason is that it is impossible to know whether the course is taught as described and just how much depth the course has. Two courses with basically the same description may have entirely different content depending upon the instructor's interests and the students' capabilities. Even the same course taught at the same university may vary from year to year depending upon which professor teaches it and what text is used. The purpose of assessing summative learning directly is to determine what computer science majors actually know, not what college catalogs claim to teach.

Other Dimensions of Learning in Computer Science

Thus far we have focussed only on the question of topics that constitute computer science. To define the domain of student learning in computer science, and to attempt to assess that learning, we must be concerned with at least three other questions about dimensions of learning:

1. Should a test focus entirely on what is taught, or is it appropriate to include topics that might be or should be taught?

2. Should it query the student on specific facts or on general principles or both?

3. Are there learned abilities that a student must acquire besides those for which there is a course title, and can we assess those abilities?

In the process of test development, the importance of distinguishing between what is learned and what ought to be learned is probably overrated. Essentially, we try to assess how much students actually know compared with how much we think they ought to know. The reason for making this point is that committees can spend a considerable amount of time debating the differences. It may be more productive to define test content in terms of what we expect at least some students to have learned. Then the question of whether they learned it or simply "ought to have learned it" becomes an empirical question settled by administering the test.

With this approach to measuring student learning, the debated question of whether calculus should be required of computer science majors and whether it should be included in a summative measure of computer science reduces to the answerable question of whether any students do learn calculus in their computer science program. The answer to that question is clearly yes, and since some do learn it, it should be measured.

The second question is whether the purpose of a test is to assess understanding of general principles, or to measure specific factual knowledge. The GRE Computer Science Test was designed to predict performance in computer science at the graduate level and therefore focuses more on understanding principles of computer systems, for example, than on knowledge of a specific system. A test of summative learning in computer science must cover general principles, but it may also measure specific knowledge, such as particular programming languages. It may be useful, for example, to determine what proportion of graduates in computer science have learned PASCAL compared with other programming languages.

In the area of programming languages, for example, we could assess students' understanding of basic principles--such as control structures and data types--principles that would apply to any higher level language. In addition, we may want to test their knowledge of specific programming languages such as C, Cobol, Fortran, or Ada. In the area of operating systems, we could assess students' understanding of operating system principles, and we might also want to determine how much they know about specific systems such as UNIX.

Virtually every topic in computer science can be broken down in this manner, with some emphasis on abstract principles and some emphasis on specific knowledge. One of the differences we would expect to find between colleges is in their relative emphasis on basic principles versus concrete knowledge. If that is the case, we would have to include test items at both extremes of this dimension.

Critical Abilities Directly Related to Computer Science

When we think of summative learning in a discipline, we think of more than knowledge of subject matter. We expect students to develop skills that go beyond knowledge acquisition, and this is the point of our third question. Some of the skills we might expect of a computer science graduate include the following:

- The ability to define a problem and the recognition that to solve a problem, one must be able to define it clearly.
- The use not only of "atomistic" but of "synergistic" problem-solving approaches. Most tests measure the ability to solve pieces of problems but cannot deal with an entire system of problems, interdependent variables, and relationships. It is important for a computer scientist to be able to employ synergistic, or highly complex, problem solving strategies.
- The ability to abstract, (that is to understand) the general principles that underlie a specific occurrence.
- The ability to judge what is the best solution in a particular context.
- Writing skills in a computer science context. A computer scientist must be able to communicate in writing, especially to document programs.
- Interpersonal and communication skills, including the ability to work as a team member to solve a problem, to organize and manage large software projects, and to assess and solve clients' problems.

Powers and Enright (1986) conducted a study to determine the perceptions of a sample of college faculty towards the importance of numerous analytical reasoning skills involved in graduate study. Skills that faculty in computer science departments rated as especially important can be grouped into five general categories: (1) general reasoning, (2) problem definition, (3) constructing theses or arguments, (4) analyzing arguments, and (5) the avoidance of specific kinds of reasoning errors or fallacies.

The last of these reasoning skills, the identification and avoidance of reasoning errors, is of particular interest in our approach to assessment. What it means is that computer science majors should have learned not to make the following types of errors:

- Applying a formula, algorithm, or other rule without sufficient justification.
- Relying solely on narrative or description in papers and reports when analysis is appropriate.
- Searching for a complicated solution when an obvious simple one exists.
- Being unable to generate hypotheses independently.

There is no real agreement among cognitive psychologists as to whether these reasoning abilities exist as pure constructs, or whether they are meaningful only in a specific context. For example, "being able to identify more than one approach to solving a problem" is an ability that is desirable for a computer scientist to have. But we do know that not everyone who has this ability applies it to all kinds of problem solving, just as we know that a "creative" person can be creative in one context (e.g., painting) and not at all creative in another context (e.g., thinking of ways to earn a living). In the development of an indicator to determine a student's strength in this ability, we would attempt to measure the "ability to identify more than one approach to solving a problem in the context of computer science." We would construct all questions or problems in a computer science context and not attempt to measure any ability that is expressed abstractly.

In this section we have attempted to show that summative learning begins with the knowledge described by the topics contained in an agreed-upon curriculum. It must go further, however, to include specialized topics plus critical skills and abilities that may be more difficult to define and measure.

Developing and Validating Computer-Based Indicators of Student Learning in Computer Science

Where We Are Now

From our review of the content domain of computer science, we developed a detailed topic outline that included the common core as well as many specialty areas. We found the procedure to be similar to conventional test development, though it was easier because those content areas that committees normally spend time debating could be placed in the category of specialty areas, that is, those topics that are covered in some degree programs and not in others. Furthermore, we concluded that there are many complex thinking and reasoning skills that students in computer science must develop, and those skills should also be assessed.

Our review of the current state-of-the-art in computer-interactive testing (CIT) indicated that applications of CIT are growing very rapidly, and not only is a considerable effort going into the development of expert systems, but their applications are becoming practical and affordable. The practicality of designing computer-based indicators depends to a large extent on whether CIT can satisfy the conditions required for an indicator of student learning, as outlined in this report. Our first step, therefore, is to review those conditions in the light of CIT.

An indicator must be an indicator of something. The linkage between the indicator and what it indicates must be clearly evident. This is a matter of establishing the indicator's validity. Whether the indicator is based on conventional test scores or CIT, it must be validated. If a validated paper-and-pencil test is converted to an adaptive test, or even if it is simply delivered by computer in its existing form, it must be validated again because changing the method of presentation can affect validity.

Validating an indicator involves validating the test (CIT or otherwise) at the two or more points in time that it will be administered to measure growth. If the same test is to be used to measure student knowledge upon admission and at graduation, the test has to be validated on incoming freshmen and on graduating seniors. We cannot assume that an item measures the same skills at each point in time. As an example, consider an algebra "word" problem. It may be worked easily by someone who knows algebra and can simply formalize it and compute the answer. Someone who has not studied algebra may use trial-and-error methods of guessing and testing solutions. The person may even invent enough algebra to do the problem. That makes it a different problem, requiring (and measuring) different skills.

If we design a CIT battery to diagnose the strengths and weaknesses, or relative emphases, of computer science programs in different institutions, we will need a test producing diagnostic subscores. Part of the validation task will be to demonstrate that indicators of student learning in each of the diagnostic areas are in fact indicators of different learning. Suppose an indicator of student learning about computer hardware always showed a gain (or loss) when an indicator of learning about software showed a gain (or loss). In addition, suppose the indicators showed that institutions that were strong in hardware were also strong in software, and vice versa. We would begin to suspect that we had a single composite indicator, not two different indicators.

Demonstrating that two indicators are representing different realities is the task of construct validity. In general, construct validity requires not only that a test be measuring what it purports to measure, but, in addition, that it not be measuring something else (Cronbach, 1971, and Campbell and Fiske, 1959). The way this is generally studied is by factor analysis. In the example cited above, there should be a factor for the construct we call "knowledge of hardware" and another factor for the construct we call "knowledge of software." Those constructs, while they may be correlated, are not the same and should be represented by distinctly different factors. A statistical procedure for confirming that the "factor structure" of a test is consistent with its design is called confirmatory factor analysis. Examples of studies employing confirmatory factor analysis to study construct validity are available for the SAT (Rock and Werts, 1979 and Dorans and Lawrence, 1987), the GRE General Exam (Rock, Werts, and Grandy, 1982), and the New Jersey Basic Skills Test (Grandy, 1980).

Establishing that there are linkages, therefore, between indicators and what they purport to indicate goes beyond content analysis. It involves a rigorous statistical process, not simply inspection of test content.

An indicator of summative learning is actually a composite indicator of a more broadly defined construct. All of the subtopics of a test, weighted somehow and added together, form this indicator. The rationale for weighting topics is rather arbitrary, but tests have always consisted of a sampling of items covering different aspects of a subject, and that sampling is based on expert decisions regarding the relative proportions of topics covered. The decision of a test development committee to have twice as many items on software as on theory is one of expert opinion about the relative importance of the two areas of learning. Any indicator of summative learning is less informative than multiple indicators of learning for specific topics. Ideally, we should have both. Multiple indicators can be added together (weighted as desired) and summarized in a single statement to give a summative indicator for whatever purpose it is required.

Developing indicators of higher order thinking skills is not so much a measurement problem as a conceptual one. "Higher order thinking skills" is a broadly defined construct which has to be broken down into more narrowly defined constructs and then linked with manifestations (or indicators) of those constructs. If this can be done, CIT is likely to be more successful than paper- and-pencil tests at assessing the quality of these skills. The computer, for example, may be able to present a complex problem and evaluate the quality of the examinee's solution. We will discuss this capability later in this section.

Two categories of indicators that we defined earlier in this paper were descriptive and quantitative. CIT is well suited to producing both kinds. An example of a quantitative statistical indicator is an increase in a mean test score over a 4-year period. To measure this growth with a paper-and-pencil test, we would administer the test to incoming freshmen, compute a score, administer the same test or a parallel form upon graduation, and compute a second score. Growth would consist of the difference between the two scores. Using a computerized adaptive test, the student would take easy items upon admission to college, harder items (hopefully) upon graduation, and because the items had been calibrated when the test was developed, the student's growth could be represented in a score similar to the score used for the paper-and-pencil test. The advantage to the adaptive test would be that it would be shorter and it would present the student only with items appropriate to his/her level of knowledge.

A descriptive statistical indicator can be produced by a mastery (or minimal competency) examination. Upon admission, for example, students take a short test to see if they already know how to program in PASCAL. A standard has already been set so that if a student passes the test, he/she does not have to take the introductory PASCAL course. Students who fail must take the PASCAL course and pass the final test before going on to take more difficult programming courses. The percent who pass the first time is a descriptive statistical indicator of the percentage of students who know PASCAL when they enter. The percentage who pass at the end of the course is a descriptive statistical indicator of student learning of PASCAL in the course. The advantage of using CIT is that for most students it is shorter than a paper-and-pencil test, and students will be given different items at the end of the course than at the beginning, even though it is the "same" test.

In summary, it appears that CIT has the potential to produce better indicators of student knowledge and student learning than do paper-and-pencil tests. Tests are likely to be shorter, and their psychometric properties (scaling, etc.) satisfy the needs of many kinds of measurement, whether for individuals or groups, and whether descriptive or quantitative. CIT even offers some hope of measuring complex thinking skills, if we are able to define those skills clearly and specifically.

Where We Are Going

One can envision a time when, by the push of a magic button, a computer produces a record of the state of knowledge, skills, and abilities of every computer science major in the nation. By asking the megacomputer the right questions, it computes instantly whatever indicators you could possibly want. We are in the year 2088, and The Machine produces the following indicator information:

1. The programming styles of 80 percent of the sophomores in computer science at Alpha College are so poor that the programs are unreadable. By the time they are seniors, however, 98 percent are writing readable programs.

2. At Omega College, the computer science department has set standards whereby seniors have to demonstrate mastery of 24 subject areas before they can graduate. This year, 6 students failed to pass the area 5 exam, possibly because they used a new instructional program. We will check other colleges using that program to see if they are having similar problems.

3. Seniors at College Tau showed an average score increase of 455 points since their freshman year on the Basic Core Computer Science Test. This places them in the 65th percentile on Basic Computer Learning.

4. A new programming language called QUAKQUAK is now used exclusively by students in 93 percent of all 4-year colleges, whereas 2 years ago, only 27 percent of the students in those colleges could even recognize a program written in QUAKQUAK.

5. Between their freshman and senior years, students at Cerebral College succeed in mastering, on the average, 17 out of the 20 Universal Computer Science Topics. Compared with the rest of the nation, this puts Cerebral College in the top ten. The College is still weak in Topic 19 and has a position open to hire a Topic-19 specialist.

With the ideal indicator generator, this list could go on indefinitely. Unfortunately we have no such machine. It is a useful heuristic device, however, to have in mind a futuristic real-time machine that could provide instantaneous measurements of student knowledge, skills, and abilities, and could generate indicators ad infinitum. We could access it whenever we needed any information for educational reform, policy making, deciding which college to attend, or which graduates to hire.

How We Can Get There

At the present time, we can shorten testing time with adaptive tests, and we can diagnose errors in more complex thinking skills with the use of expert systems. Even with the shortest adaptive tests, however, we cannot measure knowledge of each and every topic in computer science, much less measure related higher-order thinking skills, in less than many hours, days, or weeks of testing. There are, nonetheless, ways that we can begin to develop indicators of student learning in computer science that employ CIT and that will be superior to paper-and-pencil testing.

A reasonable first step is to take an existing computer science test and convert it to an adaptive test. A suitable prototype would be the new Major Field Achievement Test (MAT) in Computer Science that has just been introduced by the GRE Board and ETS. This test was designed for outcomes assessment and is based on the same specifications as the GRE Subject Test in Computer Science which our advisory committee reviewed for this project and agreed was suitable in content.

The test specifications (appendix B) are divided into five broad content areas, the fifth of which is "Special Topics." The test, therefore, actually contains only four defined content areas: (1) Software Systems and Methodology, (2) Computer Organization and Architecture, (3) Theory, and (4) Computational Mathematics. Once the Computer Science MAT has been administered to a large enough number of examinees, items can be calibrated, and we can see whether they are suitable for development into an adaptive test. This will take awhile because a large pool of items in each of the four content areas will have to be calibrated, and a single form of the test contains too few items. There may have to be 8 or 10 forms of the test given before there are enough calibrated items in the item pool to produce an adaptive test. Once those forms have been administered, it may be possible to develop an adaptive test in computer science that is diagnostic in the four broad areas defined by the specifications.

If this project were successful, it would form a cornerstone for the design of a larger system of computer-based indicators. Consider the implications of having accomplished just this first step. Institutions that are relatively strong or weak in their mathematics emphasis, for example, could identify themselves. If the test is sufficiently reliable at the individual level, graduates would learn how well they compare with other graduates and may wish to submit their scores to a potential employer. If there are enough easy items in the test, it could be administered to incoming freshmen intending to major in computer science, and their gains in knowledge by the time they graduate could be determined. If there are not enough easy items to make the test suitable for freshmen, those items could be written and calibrated as well. When calibrated items exist that cover the entire range of abilities in each of the four content areas, the test could be used to produce four indicators of student learning in computer science, one corresponding to each topic in the specifications. Summative learning would be reported as the sum of the four scores, with each one weighted in accordance with the predetermined weighting contained in the current specifications.

With that adaptive test as a foundation, we could expand on the system by refining each content area into more specific knowledge areas. "Theory," for example, now consists of three areas: automata and language theory, correctness of programs, and analysis of algorithms. Depending on the time taken to complete the test, it may be possible to produce more detailed diagnostics. It is premature to speculate as to how many subtests could be administered within a reasonable time.

Another direction that we could take is to assign a programming problem and to develop a scoring system. It seems most appropriate to require computer science majors to write and debug a program successfully, and the requirement is certainly not a new one. Carnegie-Mellon University has been using an on-line competency test in their introductory computing courses for about six years (Carrasquel, Goldenson, and Miller, 1985; and Stehik and Miller, 1985). In order to pass any introductory computer science course offered to undergraduates, a student must pass the Mastery Examination. The exam attempts to simulate a "real world" problem-solving environment. Students are given non-trivial problems to solve, and the authors argue that by using traditional testing methods it is extremely difficult to assess adequately the kinds of skills being measured by the Mastery Examination.

For each student, the Mastery Examination program selects at random one of nine problems, and the student is given a five-hour block of time in a secure computing environment to write the program. During the exam, students have access to a terminal, a directory, scratch paper and pencil, all of the necessary input files, format sheets with example input and output formats, help sheets for some exams, a variety of Pascal books, and access to all relevant help files. Prior to the exam, students have access to actual Mastery Exam questions and other practice material.

In order to be comparable, each exam problem is designed to fit the same schema. First, students must decide on a data structure. Second, they must all write similar routines regardless of the specific question that is drawn. To obtain a minimal passing grade, they must write a program that perfoms both simple I/O and sequential search. They obtain higher grades by successfully modifying records, modifying the data structure by inserting and deleting records, and by sorting the structure.

Students then record their entire interactive session in a separate file and then execute their program within a program called "photo." By following a prescribed test script, they demonstate everything that the program is capable of doing. Once the students demonstrate their program, they print all of their files to be graded. The actual grading is not done by computer.

The AP Computer Science test has a much simpler programming assignment, and as we discussed earlier, the student's program can be scored fairly successfully using MicroPROUST. For seniors in computer science, we would assign a more difficult programming task. Considerable research may be necessary to design a scoring program for a very difficult assignment.

There is one problem that CIT experts frequently raise in conjunction with the development of expert systems to score constructed response questions: there is no well defined way to decide what kinds of errors are more serious than others, and consequently, there is no basis for weighting errors and producing justifiable numerical scores. A reply to this concern is that we do not always need numerical scores. The assumption that we must decide what kind of error is more serious than another kind of error is simply not tenable. A description of the kinds of errors made is far more informative than a score. Indicators of student learning will be more useful if they report that students make fewer errors of type X by the time they are seniors than they did as sophomores, but they make more errors of type Y. To say that the seriousness of their errors is greater (or less) now that they are seniors merely passes judgment.

It does not give us the information we need for remediation. This example illustrates the usefulness of descriptive indicators as opposed to quantitative ones. Scores do not have to be the outcome of testing.

The use of expert systems for evaluating constructed response questions need not be restricted to programming assignments. In any discipline, experts argue that there are important kinds of knowledge, skills, and abilities that cannot be measured by multiple-choice test items. Rather than attempting to define and measure higher-order thinking skills, the advisory committee for this project suggested seven problem areas that they believe should be tested but which cannot be tested with multiple-choice items. For some of these areas, they designed test items to exemplify the problems and ways they would evaluate student solutions. The designated appendices contain the sample test items. Some sample test items exemplify more than one of the seven problem areas.

1. Problems requiring a computer for their solution because of their size and/or complexity, or because a grader would require a computer to check the correctness of the solution.

In mathematics, the solution to a large matrix problem would require a computer to solve. More importantly, however, there are problems that are impossible or impractical for a human grader to score. Appendix D presents two examples from computer science. The first is drawn from theory of computation; the second is from logic design.

2. Problem-solving by successive querying requires that the examinee seek additional information in order to solve the problem. Because the problem is not completely specified, the examinee must know where to search for information to solve the problem.

Circuit debugging is a fine example of a problem that demands successive querying. Many expert systems of this type are in use for training technicians in troubleshooting. The Logic Gate Tutor mentioned earlier in this report is an example. The computer presents a schematic diagram of a circuit and informs the technician that the circuit is defective. It is up to the technician to isolate the defective gate. Using a mouse, the technician can trace the circuit and take a reading at any gate or replace the gate if it is suspect. The program provides feedback to the technician. If he or she is successful in debugging the circuit, the technician is exited from the system. If not, the program presents a series of questions to isolate the technician's conceptual problems. It does this by evaluating the techician's error patterns using error tables built into the program's knowledge base. The error patterns may indicate that the technician is confusing two types of gates. If so, the program can provide appropriate instruction.

What is important about this type of problem is that it not only tests the examinee's ability to recognize logic gates and understand their function (which could probably be tested with muliple-choice items), but it requires that the technician know what questions to ask. It is this aspect of the problem that puts it in a real-world context.

3. A problem in which the solution is difficult to generate but trivial to verify.

Problems of this sort are common in mathematics. If an examinee is asked to solve a set of simultaneous equations and is provided with five multiple-choice answers, the solution is easy to identify by successively substituting the values given in each answer choice until one set fits.

Appendix E gives an example from computer science.

4. Non-textual problem presentation. There are many circumstances in which a written, or textual, presentation is not the best way to present a problem. Often graphics, a video segment, or sound is not only more appropriate but essential. The Logic Gate Tutor described earlier is such an example. As the examinee points to a gate to test or replace, the picture changes, presenting the information requested. This could not be done with a drawing in a test booklet. Similarly, the U.S. Navy developed a simulation called "Steamer" to train shipboard personnel in the operation of steam engines for large ships. It uses elaborate advanced graphics to show what happens in the system when a component fails. This simply could not be done on paper.

The applications of non-textual presentation are too numerous to attempt to list. Testing an examinee's ability to observe essential action in an interpersonal situation must be done with video. Medical technicians must be able to see an injured patient (or video version of a patient) in order to know what symptoms to look for prior to taking action. Many aspects of musical performance and listening require actual music; a computer would enable the examinee to manipulate the music.

These testing applications apply to disciplines other than computer science. Computer scientists, however, must be knowledgeable about graphics, interactive videodisk technology, computer music, etc. The best way to test their knowledge and skills would be with a hands-on, real-world problem.

5. Stepwise refinement problems. Real-world problems are generally complex and often not clearly specified. A single right answer probably does not exist, but there are many reasonably acceptable solutions that can be refined as time goes on. Stepwise refinement refers to the process of breaking down a large problem into smaller, workable ones, and finding alternatives when the smaller problems change or turn out to be unsolvable. The ability to analyze a problem in this manner requires special skills. In its simplest form, it requires a person to decompose a problem into small, workable ones. But in most real-life situations, problem-solving is not so well defined.

Consider as an example, a typical work team. The leader has the total problem in mind and breaks it down so that subordinates can work on pieces of the problem. Often the subordinates, if given the entire problem to break down, would not know where to begin. Suppose the leader breaks down the problem and assigns concrete sequential tasks to a subordinate. When something does not work as expected, the subordinate is back with the question, "Now what do we do?"

Stepwise refinement problems resemble troubleshooting tasks in the sense that the problem is not completely specified, and the action taken at each step often depends on the outcome of the previous action. But there are major differences. Troubleshooting has a solution (the machine has to run) and a finite number of possible paths to that solution (there are just so many components that can be replaced). In a stepwise refinement problem, there may be no solution, and the paths have not yet been laid. Developing indicators of summative learning in computer science is a stepwise refinement problem.

It is clear that a paper-and-pencil test cannot present a real-life stepwise refinement problem. There is some possibility that a computer can. Designing a computer system itself is such a problem. It may be possible to develop an expert system examination that can trace and evaluate this kind of thinking.

6. Problem solving with the aid of hints. There are many times when an examinee can solve part of a problem and, with the aid of a hint, could complete the rest of the problem. Often instructors believe that a student should receive partial credit under these conditions.

A test delivered by an expert system can "follow" an examinee's solution steps and give hints if the examinee requests them. Appendix F offers four programming examples where hints are used.

7. Problems with multiple correct solutions. Most real-world problems have more than one solution. Writing computer programs, writing essays, and producing creative arts are all areas in which the quality of a solution, or the mere existence of any solution, is the criterion of success.

All of the problems in the appendices allow for many solutions. A multiple-choice test, by giving solutions, clearly cannot measure the examinee's ability to create solutions.

The seven problem types discussed here are examples of the many types of testing that are possible, and in some instances, are already in use. So far, we have discussed ways that these problem-solving skills can be assessed. We have not addressed the development of indicators of these skills. At this point in time, it seems that defining the skills and developing some questions that measure them will be a major accomplishment. The task of representing the results as indicators should be trivial by comparison.

Many of the skills that we wish to assess may be inappropriate to express quantitatively. Information from the Logic Gate Tutor, we recall, identified examinees' misconceptions. The great advantage of expert systems may be that they can produce decriptive diagnoses of weaknesses and point out areas of strength. When we look again at the indicator statements produced by our futuristic indicator generator, we see that the most meaningful statements are based on the results of mastery examinations rather than on scaled test scores.

Perhaps if we develop more intelligent tutors containing diagnostic tests (like the Logic Gate Tutor) in addition to mastery tests using hierarchies of testlets (like the NCARB test), we can build them into the regular instructional program. The student learns as the tutor diagnoses errors and presents instruction. The diagnosed errors are a real-time indicator of that student's knowledge at the moment. This idea has been proposed by Bunderson, Inouye, and Olsen (in press). Its realization would be not far from our futuristic indicator generator.

References

Anastasi, A. (1988). Psychological Testing, 6th ed. New York: Macmillan.

Anderson, J. R. (1983). The Architecture of Cognition. Cambridge, MA: Harvard University Press.

Anderson, J.R. and Reiser, B.J. (1985). "The LISP tutor." Byte, 10, 159-179.

Anderson, J. R. and Skwarecki, E. (1986). "The automated tutoring of introductory computer programming." Communications of the ACM, 29, 842-849.

Association for Computing Machinery. (1983). ACM Recommendations for Computer Science (volume 1). New York, NY: Association for Computing Machinery.

Bennett, R.E., Gong, B., Kershaw, R.C., Rock, D.A., Soloway, E., and Macalalad, A. (in preparation). Agreement between Expert System and Human Ratings of Constructed-Responses to Computer Science Problems. Princeton, NJ: Educational Testing Service.

Bunderson, C.V., Inouye, D.K., and Olsen, J.B. (In press). "The four generations of computerized educational measurement." In R. Lynn (Ed.), Educational Measurement, 3rd edition.

Campbell, D., and Fiske, D. (1959). "Convergent and discriminant validation by the multitrait-multimethod matrix." Psychological Bulletin, 56, 81-105.

Carrasquel, J., Goldenson, D. R., and Miller, P. L. (1985). "Competency testing in introductory computer science: the Mastery Examination at Carnegie-Mellon University." In Proceedings of the 1985 Computer Science Conference. ACM-SIGCSE. New Orleans.

College Entrance Examination Board. (1984). CLEP Test Information Guide--Computers and Data Processing. New York, NY: College Entrance Examination Board.

College Entrance Examination Board. (1985). Computerized Placement Tests: A Revolution in Testing Instruments. New York, NY: College Entrance Examination Board.

College Entrance Examination Board. (1986). The Entire 1984 AP Computer Science Examination and Key. New York, NY: College Entrance Examination Board.

College Entrance Examination Board. (1987). Advanced Placement Course Description: Computer Science. New York, NY: College Entrance Examination Board.

Cronbach, L. J. (1971). "Test validation." In R. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education.

Dorans, N.J. and Lawrence, I.M. (1987). The Internal Construct Validity of the SAT. RR-87-35. Princeton, N.J.: Educational Testing Service.

Educational Testing Service. (1983). ETS Standards for Quality and Fairness, Princeton, NJ: Educational Testing Service.

Educational Testing Service. (1988). "Breakthrough development in computerized testing offers shorter tests, more precise pass-fail decisions." ETS Developments, 33, 3-4.

Eignor, D. R. and Cook, L. L. (1983). "An investigation of the feasibility of using item response theory in the preequating of aptitude tests." Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Gialluca, K.A. and Weiss, D.J. (1979). "Efficiency of an adaptive intersubtest branching strategy in the measurement of classroom achievement." Research Report 79-6. Minneapolis: Computerized Adaptive Testing Laboratory, University of Minnesota.

Gialluca, K.A. and Weiss, D.J. (1981). "Dimensionality of measured achievement over time." Research Report 81-5. Minneapolis: Computerized Adaptive Testing Laboratory, University of Minnesota.

Gibbs, N. and Tucker, A. (1984). "On accreditation." Communications of the ACM, 27, 5, 411-412.

Gibbs, N. and Tucker, A. (1986). "A model curriculum for a liberal arts degree in computer science." Communications of the ACM, 29, 3, 202-210.

Gill, T. A. and Stehlik, M. J. (1985). Grading the Advanced Placement Examination in Computer Science. New York, NY: College Entrance Examination Board.

Gitomer, D.H. (1987). "Using error analysis to develop diagnostic instruction." Paper presented at the Military Testing Association, Ottawa.

Gong, B. and Sleeman, D. (1985). Remediation of misconceptions of high school BASIC programmers through individualized tutoring. Final report to the Spencer Foundation.

Graduate Record Examinations Board. (1986). Practicing to take the GRE Computer Science Test. Princeton, NJ: Educational Testing Service.

Grandy, J. (1980). Analysis of the Subscale Structure of Test Batteries: A Confirmatory Study of the Interrelationships of CGP and NJ Basic Skills Test Scores. RR 80-25. Princeton, NJ: Educational Testing Service.

Green, B. F. (1986). "Construct validity of computer-based tests." In H. Wainer and H. Braun (Eds.), Test Validity. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K. (1968). "Effects of item order and anxiety on test performance and stress." Paper presented to Division D, American Educational Research Association annual meeting, Chicago, February.

Johnson, W. L., Soloway, E., Cutler, B., & Draper, S. (1983). A Bug Catalogue: I. Report No. 9, Cognition and Programming Project, Department of Computer Science, Yale University.

Johnson, W. L. and Soloway, E. (1984). "Intention-based diagnosis of programming errors." Proceedings of the National Conference on Artificial Intelligence. Ausin, TX.

Johnson, W. L. (1985). Intention based diagnosis of errors in novice programs. Report No. 23, Cognition and Programming Project, Yale University.

Kingston, N. M. and Dorans, N. J. (1984). "Item location effects and their implications for IRT equating and adaptive testing." Applied Psychological Measurement, 8, 146-154.

Lockheed, M. E. and Mandinach, E. B. (1986). "Trends in educational computing: decreasing interest and the changing focus of instruction." Educational Researcher, May, 1986.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. New York: Lawrence Erlbaum Associates.

MacNicol, K. (1956). "Effects of varying order of item difficulty in an unspeeded verbal test." Unpublished manuscript. Princeton, NJ: Educational Testing Service.

Maurelli, V.A. and Weiss, D.J. (1981). "Factors influencing the psychometric characteristics of an adaptive testing strategy for test batteries." Research Report 81-4. Minneapolis: Computerized Adaptive Testing Laboratory, University of Minnesota.

Mollenkopf, W. G. (1950). "An experimental study of the effects on item analysis data of changing item placement and test-time limit." Psychometrika, 15, 291-315.

Monk, J. J. and Stallings, W. M. (1970). "Effect of item order on test scores." Journal of Educational Research, 63, 463-465.

Montag, M. et al. (1984). Standardized Test of Computer Literacy. Available through Instructional Resources Center, Iowa State Univ.

Mulder, M. C. and Dalphin, J. (1984). "Computer science program requirements and accreditation." Communications of the ACM, 27, 330-335.

Oakes, J. (1986). Educational Indicators: A Guide for Policymakers. OPE-01 (October 1986) Center for Policy Research in Education. Washington, D.C.: U.S. Department of Education, Office of Educational Research and Improvement.

Oltman, P.K. (1982). Content Representativeness of the Graduate Record Examinations Advanced Tests in Chemistry, Computer Science, and Education (GRE Report GREB No. 81-12P). Princeton, NJ: Educational Testing Service.

Poplin, M. S. et al. (1984). "Computer Aptitude, Literacy, and Interest Profile." Available from PRO-ED, 5341 Industrial Oaks Blvd, Austin, TX.

Powers, D.E. and Enright, M.K. (1986). Analytical Reasoning Skills Involved in Graduate Study: Perceptions of Faculty in Six Fields. GRE Board Professional Report No. 83-23P. Princeton, NJ: Educational Testing Service.

Psychological Corporation. (1986). Computerized Adaptive Differential Aptitude Test. San Antonio, TX: Psychological Corporation.

Putnam, R., Sleeman, D., Baxter, J., and Kuspa, L. (1984). A Summary of Misconceptions of High School Basic Programmers. Occasional report #010. Technology Panel, Study of Stanford and the Schools, Stanford University, June 1984.

Ralston, A. and Shaw, M. (1980). "Curriculum '78--Is computer science really that unmathematical?" Communications of the ACM, 23, 67-70.

Reiser, B. J., Anderson, J. R., and Farrell, R. G. (1985). "Dynamic student modelling in an intelligent tutor for LISP programming." Paper presented at the Proceedings of the International Joint Conference on Artificial Intelligence, Los Angeles.

Rock, D. A. and Werts, C. E. (1979). Construct Validity of the SAT across Populations--An Empirical Confirmatory Study. RDR 78-79, No. 5. Princeton, NJ: Educational Testing Service.

Rock, D.A., Werts, C.E., and Grandy, J. (1982). Construct Validity of the GRE Aptitude Test across Populations--An Empirical Confirmatory Study. GREB No. 78-1P. Princeton, NJ: Educational Testing Service.

Sax, G. and Carr, A. (1962). "An investigation of response sets on altered parallel forms." Educational and Psychological Measurement, 22, 371-376.

Shaw, M. (1984). The Carnegie-Mellon Curriculum for Undergraduate Computer Science. New York: Springer-Verlag.

Sleeman, D. and Brown, J. S. (1982). Intelligent Tutoring Systems. New York: Academic Press.

Sleeman, D., Baxter, J., Putnam, R., & Kuspa, L. (in press). Misconceptions in high school Pascal programmers. Preliminary version available as "Pascal and high school students: a study of misconceptions." Occasional Report #009. Technology Panel, Study of Stanford and the Schools, Stanford University, August 1984.

Spohrer, J., Pope, E., Lipman, M., Sack, W., Freiman, S., Littman, D., Johnson, L., and Soloway, E. (1985). Bug Catalogue: II, III, IV. Report No. 24, Cognition and Programming Project, Yale University.

Stehlik, M. J. and Miller, P. L. (1985). Implementing a Mastery Examination in Computer Science. CMU-CS-85-175, Pittsburgh, PA: Carnegie-Mellon University.

Stern, J.D. (1986). The Education Indicators Project in the U.S. Department of Education. (April 1986). Washington, D.C.: Center for Statistics, U.S. Department of Education.

Wainer, H. (1983). "On item response theory and computerized adaptive tests." The Journal of College Admissions, 28 (4), 9-16.

Wainer, H. and Kiely, G. L. (1987). "Item clusters and computerized adaptive testing: a case for testlets." Journal of Educational Measurement, 24, in press.

Waters, B.K. (1981). The Test Score Decline: A Review and Annotated Bibliography. Technical Memorandum 81-2. Washington, D.C.: Directorate, Office of the Secretary of Defense.

Ward, W. C., Kline, R. G., and Flaugher, J. (1985). College Board Computerized Placement Tests: Summary of Pilot Testing Results. Princeton, NJ: Educational Testing Service.

Werts, C. E. and Linn, R. L. (1970). "Path analysis: Psychological examples." Psychological Bulletin, 74, 193-212.

Whitely, S. E. and Dawes, R. V. (1976). "The influence of test context on item difficulty." Educational and Psychological Measurement, 36, 329-337.

Yen, W. M. (1980). "The extent, causes, and importance of context effects on item parameters for two latent trait models." Journal of Educational Measurement, 17, 297-311.

Appendix A

Members of the Advisory Committee

Rafael Alonso, Assistant Professor of Computer Science, Princeton University. Dr. Alonso's background is in computer science, electrical engineering, and mathematics. He was previously employed at Bell Telephone Laboratories. His special research interests are in distributed systems, databases, operating systems, and computer architecture.

Brian Reiser, Assistant Professor of Psychology, Princeton University. Dr. Reiser is conducting research on intelligent tutoring systems, problem solving, and memory.

Louis Steinberg, Associate Professor of Computer Science, Rutgers University. Dr. Steinberg's research interests lie in expert systems and machine learning.

Kenneth Supowit, Assistant Professor of Electrical Engineering and Computer Science, Princeton University. Dr. Supowit has taught courses in very large scale integrated circuit (VLSI) design and in analysis of algorithms. He has published extensively and has recently received a National Science Foundation Presidential Young Investigator Award. Dr. Supowit's research interests lie in the areas of computer-aided design of digital systems, computer architecture, and combinatorial algorithms.

Appendix B

Detailed Specifications: GRE Advanced Test in Computer Science

I. Software Systems and Methodology

A. Data organization

1. Abstract data types (e.g., stacks, queues, lists, strings, trees, sets)
2. Implementations of data types (e.g., pointers, hashing, encoding, packing, address arithmetic)
3. File organization (e.g., sequential, indexed, multilevel)
4. Data models (e.g., hierarchical, relational, network)

B. Organization of program control

1. Iteration and recursion
2. Functions, procedures, and exception handlers
3. Concurrent processes, interprocess communication, and synchronization

C. Programming languages and notation

1. Applicative versus procedural languages
2. Control and data structure
3. Scope, extent, and binding
4. Parameter passing
5. Expression evaluation

D. Design and development

1. Program specification
2. Development methodologies
3. Development tools

E. Systems

1. Examples (e.g., compilers, operating systems)
2. Performance models
3. Resource managment (e.g., scheduling, storage allocation)
4. Protection and security

II. Computer Organization and Architecture

A. Logic design

1. Implementation of combinational and sequential circuits
2. Functional properties of digital integrated circuits

B. Processors and control units

1. Instruction sets, register and ALU organization
2. Control sequencing, register transfers, microprogramming, pipelining

C. Memories and their hierarchies

1. Speed, capacity, cost
2. Cache, main, secondary storage
3. Virtual memory, paging, segmentation devices

D. I/O devices and interfaces

1. Functional characterization, data rate, synchronization
2. Access mechanism, interrupts

E. Interconnection

1. Bus and switch structures
2. Network principles and protocols
3. Distributed resources

III. Theory

A. Automata and language theory

1. Regular languages (e.g., finite, automata, nondeterministic finite automata, regular expressions)
2. Context-free languages (e.g., notations for grammars, properties such as emptiness, ambiguity)
3. Special classes of context-free grammars (e.g., LL, LR, precedence)
4. Turing machines and decidability
5. Processors for formal languages, (e.g., parsers, parser generators)

B. Correctness of programs

1. Formal specifications and assertions (e.g., pre- and post-assertions, loop invariants, invariant relations of a data structure)
2. Verification techniques (e.g., predicate transformers, Hoare axioms)

C. Analysis of algorithms

1. Exact or asymptotic analysis of the best, worst, or average case of the time and space complexity of specific algorithms
2. Upper and lower bounds on the complexity of specific problems
3. NP - completeness

IV. Computational mathematics

- A. Discrete structures: Basic elements of
    1. Abstract algebra
    2. Mathematical logic, including Boolean algebra
    3. Combinatorics
    4. Graph theory
    5. Set theory
    6. Discrete probability
    7. Recurrence relations
- B. Numerical mathematics
    1. Computer arithmetic
    2. Classical numerical algorithms
    3. Linear algebra

V. Special topics

- A. Modeling and simulation
- B. Information retrieval
- C. Artificial intelligence
- D. Computer graphics
- E. Data communications

Appendix C

Computer Science Topics for Indicators of Summative and Specialized Learning

A. Common Core of Learning in Computer Science

1. Programming

1.1 Algorithms
- Concept and properties
- Role in problem-solving process
- Constructs and languages to facilitate expression

1.2 Programming languages (theory and applications)

- Subprograms
- I/O
- Language levels (machine, assembly, procedural, symbolic)
  - Applicative versus procedural
  - Compiler techniques, code generation
  - Basic syntax of higher level language
  - Regular languages (finite automata, nondeterministic finite automata, regular expressions)
  - Context-free languages (notations for grammars, properties such as emptiness, ambiguity)
  - Special classes of context-free grammers (LL, LR, precedence)
  - Control and data structure
  - Scope, extent, and binding
  - Parameter passing
  - Expression evaluation (precedence and interpretation)

1.3 Organization of Program Control

- Iteration and recursion
- Functions, procedures, and exception handlers
- Concurrent processes, interprocess communication, and synchronization
- Loop invariants, invariant relations of a data structure

1.4 Programming style

- Preparation of readable, understandable programs
- Comments
- Program docummentation
- Practical aspects of proving programs correct
- Structured programming

1.5 Debugging and verification

Use of debugging software, selection of test data
Techniques for error detection
Relation of good programming style to the use of error detection
Formal specifications and assertions (e.g., pre- and postassertion loop invariants, invariants relations of a data structure)
Verification techniques (e.g., predicate transformers, Hoare axioms)

1.6 Computational mathematics

Basic logic and elementary set theory
Elementary statistics
Analysis of roundoff error

2. Software

2.1 Computer structure and machine language

Organization of computers in terms of I/O, storage, control and processing units
Register and storage structures, instruction format and execution
Principal instruction types
Machine arithmetic
Program control
I/O operations
Interrupts

2.2 Data representation

Bits, bytes, words, and other information structures
Number representation
Representation of elementary data structures
Data transmission, error detection and correction
Fixed versus variable word lengths

2.3 Symbolic coding and assembly systems

Mnemonic operation codes
Labels
Symbolic addresses and address expressions
Literals
Extended machine operations and pseudo operations
Error flags and messages
Scanning of symbolic instructions and symbol table construction
Overall

2.4 Addressing techniques

Absolute, relative, base associative, indirect, and immediate addressing

Indexing
Memory mapping functions
Storage allocation, paging, and machine organization to facilitate modes of addressing

2.5 Macros

Definition, call, expansion of macros
Parameter handling
Conditional assembly and assembly time computation

2.6 Program segmentation and linkage

Subroutines, coroutines, and functions
Subprogram loading and linkage
Common data linkage transfer vectors
Parameter passing and binding
Overlays
Re-entrant subprogram
Stacking techniques
Linkage using page and segment tables
Terminate/Stay resident

2.7 Linkers and loaders

Separate compilation of subroutines
Incoming and outgoing symbols
Relocation
Resolving intersegment references by direct and indirect linkage

2.8 Systems and utility programs

Loaders, I/O systems, human interface with operating systems
Program libraries
Examples (e.g., compilers, operating systems)
Performance models
Resource management (e.g., scheduling, storage allocation)
Protection and security
Database management (e.g., transactions, serializability)
Distributed systems
Compilers, cross-compilers, and interpreters

3. Hardware

3.1 Computer architectures

Characteristics of, and relationships between I/O devices, processors, control units, main and auxiliary storage devices

Functional characterization, data rate, synchronization
Access mechanism, interrupts
Instruction sets, register and ALU organization
Control sequencing, register transfers, microprogramming, pipelining
Computer arithmetic

Organization of modules into a system
Multiple processor configurations and computer networks
Relationship between computer organization and software

3.2 Logic design

Basic digital circuits
AND, OR, and NOT elements
Half-adder, adder, storage, and delay elements
Encoding-decoding logic
Basic concepts of microprogramming
Logical equivalence between hardware and software
Elements of switching algebra
Combinatorial and sequential networks

3.3 Data representation and transfer

Codes, number representation
Flipflops, registers, gates

3.4 Digital arithmetic

Serial versus parallel adders
Subtraction and signed magnitude versus complemented arithmetic
Multiply/divide algorithms
Elementary speed-up techniques for arithmetic

3.5 Digital storage and accessing

Memory control
Data and address buses
Addressing and accessing methods
Memory segmentation
Data flow in multimemory and hierarchical systems
Speed, capacity, cost
Cache, main, secondary storage
Virtual memory, paging, segmentation devices
Channels

3.6 Control and I/O

Synchronous and asynchronous control
Interrupts
Modes of communication with processors

3.65 Interconnection

Bus and switch structures
Network principles and protocols
Distributed resources

3.7 Reliability

Error detection and correction, diagnostics

3.8 Parallel architectures

Single-instruction-multiple-data
Miltiple-instruction-multiple-data
Pipelined vector machines

3.9 Operating systems (Examples: UNIX, VMS, MVS, MS DOS, CPM)

4. Data Structures and File Processing

4.1 Data structures

File organization (sequential, indexed, multilevel)
Data models (hierarchical, relational, network)
Arrays, strings, stacks, queues, linked lists
Representation in memory
Algorithms for manipulating data within these structures

4.2 Sorting and searching

Algorithms for incore sorting and searching methods
Comparative efficiency of methods
Table lookup techniques
Hash coding

4.3 Trees

Basic terminology and types
Representation as binary trees
Traversal schemes
Representation in memory
Breadth-first and depth-first search techniques
Threading
Forward/backward chaining

4.4 File terminology

Record, file, blocking, database
Overall idea of database management systems
Relational/non-relational

4.5 Sequential access

Physical characteristics of appropriate storage media
Sort/merge algorithms
File manipulation techniques for updating, deleting, and inserting records

4.6 Random access

Physical characteristics of appropriate storage media
Physical representation of data structures on storage devices
Algorithms and techniques for implementing inverted lists, multilists, indexed sequential, hierarchical structures

4.7 File I/O

File control systems (directory, allocation, file control table, file security)
I/O specification statements for allocating space and cataloging files
File utility routines
Data handling (format definition, block buffering, buffer pools, compaction)

B. Special topics in computer science and related disciplines (Topics that will distinguish between colleges in their departmental emphases. These may be advanced topics, or just specific ones, like languages.)

5. Special topics

5.1 Analysis of Algorithms

Exact or asymptotic analysis of the best, worst, or average case of the time and space complexity of specific algorithms
Upper and lower bounds on the complexity of specific problems
Complexity theory, NP-completeness

5.2 Languages

Comparison of langauges
Compiler writing
Proficiency in a specific language (e.g., C, Cobol, Fortran, LISP, Pascal, ADA, Smalltalk)

5.3 Mathematics for computer science

Probability and statistics
Combinatorial analysis
Abstract algebra
Graph theory
Linear algebra
Mathematical logic
Set theory
Recurrence relations
Order notation
Matrix operation (e.g., Gaussian elimination)
Nonlinear equations
Numerical analysis
Calculus
Ordinary differential equations
Queueing theory

5.4 Computer engineering (hardware)

Digital design
Advanced computer architecture (e.g., fault tolerant systems, parallel machines, data base machines)
Microprocessor-based systems design
Telecommunications
Networks
VLSI design

5.5 Software engineering

Programming environments
Human factors
Software reusability
Programming tools (e.g., LEX, Lint, YACC, language-specific editors)

5.6 Graphics

Graphics hardware
Fundamental graphics operations (scan-conversion, clipping)
Graphics packages
Stills and animation
Hidden surface elimination
Anti-aliasing techniques

5.7 Artificial intelligence and robotics

Problem solving
Learning

Machine vision
Expert systems
Robotics
Natural language processing

5.8 Related management and information science

Operations research, particularly optimization
Economics, especially project scheduling and estimation
Management, especially techniques related to automation and high-technology development
Role of computers in organizations

5.9 Related public policy

Social implications of large-scale computing
Consumer issues in personal computers
Policy issues arising from computing and communications
Computer models for policy analysis
Legal issues such as ownership of software, liability, security
International standards (ISO, ANSI, IFIPS)

Appendix D

Two Problems with Multiple Solutions that Must be Graded by a Computer

Contributed by
Assistant Professor Kenneth Supowit
Princeton University

1. Theory of Computation: Construct a finite automaton to recognize a given regular language.

> A computer (but not a human grader) could easily perform the algorithm to decide whether the automaton given by the student is equivalent to a known correct one.

2. Logic Design: Express the Boolean function f(a,b,c,d) (defined by a given truth table) as the sum of a minimum number of prime implicants.

> There may be a number of equivalent correct answers to this question. Furthermore, there may be nearly correct answers, such as logically correct formulae with slightly more than a minimum number of terms; such responses could be awarded partial credit.

Appendix E

Example of a Problem in Which the Solution is Difficult to Generate but Trivial to Verify

Contributed by
Assistant Professor Kenneth Supowit
Princeton University

1. Combinatorics: Find a closed form for a given recurrence relation.
This is a particularly good example of the shortcomings of the multiple-choice format, since it is usually trivial to verify a proposed solution to a recurrence, whereas finding the solution is the challenge. The computer must check whether the student's solution is equivalent, up to simple rules such as commutativity ($a + b = b + a$) to the solution provided by the tester. The computer might also be provided with simple combinatorial identities such as $(r^{n}) = (n.^{n}r)$ in order that it can handle a broader class of problems.

Appendix F

Example of Problem-Solving with Hints

Problems 1, 2, and 3 contributed by Assistant Professor Rafael Alonso, Princeton University.
Problem 4 contributed by Assistant Professor Kenneth Supowit, Princeton University

1. Program Generation

This question tests basic programming skills and has been used in actual midterm examination for an introductory programming course at Princeton University. The computer language used is Pascal.

Here we ask that the student fill in the body of a program that processes character data. Clearly, there are many ways to answer the question correctly. A computer can easily check that the student obtained a correct answer by running the student's program with a small amount of test data (very much in the same manner the instructor would grade the question in a paper-and-pencil exam). However, there are a number of issues that arise.

First, if a question like this one called for an efficient algorithm (for example, in writing a program that will sort an array of numbers), the computer can time the running time of the solution and determine how efficient the program is.

Secondly, while it is simple to deal with the case in which the student's solution is correct, it is less clear what the appropriate course of action should be if the program has errors. A possible approach would be to determine if by making a few simple "mutations" an incorrect program can be transformed into a correct one. But the best way to deal with an erroneous program would be to notify the student that something is wrong, and allow corrections to be made. One could permit a fixed number of debugging runs, allot a fixed amount of time for finishing, or put limits on the test-taker. Even if no limits are enforced, the time required or the number of runs needed by the student could be used to grade the student. Tools could be provided to the student to help correct the program (such as syntax-level debuggers). The test designers must also decide how much feedback to provide the student (e.g., a message such as "division by 0 in line 5" will be more helpful that a generic "incorrect program" message).

A further point to consider is that if syntactic errors (i.e., "misspellings" of the computer language) are not considered important for the understanding of a concept, students might be provided with a syntax-directed editor that will prevent such errors from being made. (As an aside, note that the student must be given time to become familiar with tools such as editors and debuggers before the exam.)

Problem 1.

We want to write a program that reads an input text file and echoes it exactly the way it is (same line structure), with one exception: Any string of repeating, contiguous characters is replaced by a single instance of that character, followed by the number of times the character was encountered (assume a maximum of 9). So for example, the text:

```
aaaaabcccddeef f f  ghh
hjkk ooo
```

produces the following output:

```
a5bc3d2e2f f f3gh2
hjk2 o3
```

(There are spaces between the f's.) The end-of-line character should be treated as a non-repeated character (not a blank), so the first character in a line is always echoed, even if it is the same as the last character in the preceding line. Blanks are to be treated like any other character.

In the blank space provided, fill in the missing body of the skeleton program on the next page (NOTE: you may put more than one Pascal statement per line of program).

```
PROGRAM compress(input, output);
{compresses repeating characters}

VAR
  c1, c2:  char;
  startofline:  Boolean;
  numseen:  integer;

BEGIN {main}

 WHILE (not eof) DO
  BEGIN
   startofline:=true;
   WHILE (not eoln) DO
    BEGIN
     IF (startofline) THEN
      BEGIN
         ************ FILL IN LINE 1 ************
         ************ FILL IN LINE 2 ************
         ************ FILL IN LINE 3 ************
      END
    ELSE
      BEGIN
         ************ FILL IN LINE 4 ************
         ************ FILL IN LINE 5 ************
         ************ FILL IN LINE 6 ************
         ************ FILL IN LINE 7 ************
         ************ FILL IN LINE 8 ************
         ************ FILL IN LINE 9 ************
      END
    END; {eoln}
         ************ FILL IN LINE 10 ***********
         ************ FILL IN LINE 11 ***********
         ************ FILL IN LINE 12 ***********
         ************ FILL IN LINE 13 ***********
    END {eof}
  END. {main}
```

line 1: ____

line 2: ____

line 3: ____

.

.

.

etc.

2. Program Prediction

To be able to develop computer programs, students must be able to predict the behaviour of non-trivial sections of code. This question asks students to hand-generate the output of a computer program. This Pascal question has also been used in a midterm examination for an introductory programming course at Princeton University.

It should be pointed out that the programs used in this kind of question can be generated randomly, providing a check against cheating. Secondly, it will be easy for the computer to check the answer given against its computed solution. In the case of an incorrect solution, pattern-matching algorithms can be used to determine how far apart the student response is from the true answer.

In case of an incorrect solution, the student may be given another chance, perhaps accompanied by a hint about where in the sequence of program output the mistake has been made (i.e., "You are missing a line of output after line 3").

Problem 2.

Simulate the following program by hand and write the output produced in the space below. Mark clearly any spaces in the output. There are not necessarily 10 lines of output.

```
PROGRAM test (input, output);
CONST
 n = 3; blank = '; dollar = '$';
TYPE
 matrix = array [1..n] of array [1..n] of char;
 fruit = (apple, banana, pear, pineapple, grape);
 digit = '0'..'9';
VAR
 ij: 1..n;
  f: friut;
  d: digit;
  m: matrix;
FUNCTION fun (a:matrix):char;
 BEGIN
      a[1,1]:='a';m[2,2]:='b';
      fun:='c';
 END;
BEGIN
 FOR f:=pear DOWNTO apple DO write( ord(f):1);
 writeln('x');
 d:='3';
 writeln(pred(succ(d)),succ(succ(d)));
 FOR i:=1 TO n DO
      FOR j :=1 TO n DO m[j,i]:=dollar;
 FOR j:=2 TO n DO
      FOR i :=j-1 TO n DO m[i,j]:=blank;
 FOR i:=1 TO n DO writeln(m[i,1],m[i,2],m[i,3]);
 writeln(fun(m));
 FOR 1:=1 TO n DO writeln(m[i,1],m[i,2],m[i,3]);
END.
```

line 1: ____

line 2: ____

line 3: ____

.

.

.

etc.

3. Parsing

One of the skills learned in a modern compiler course is that of using a tool to automatically generate the parsing phase of a compiler. (Parsing is the process of recognizing the keywords, expressions, etc. of a program to determine if it is a legal program for a given programming language.) The process of constructing a parser for a language can be completely automated, and all compiler writers use such tools routinely.

These tools typically take as input a grammar in some standard format, and output parsing code, that is a program segment in a high-level language that can be used as part of a compiler to parse arbitrary programs. However, their use is not completely straightforward, since they only work without complaint if the input grammer contains no inconsistencies (i.e., "shift-reduce" conflicts). As the student gets feedback from the program that his input grammar is ambiguous, he or she is required to modify the description of the grammar (while preserving its essential form) so conflicts will be eliminated.

Clearly, in order to test whether students know how to use such a tool, a computer must be made available to them so that they can show how they would generate a parser given a test grammar. Furthermore, the feedback mechanism described above requires that students be able to interact with a computer to complete the modification of the grammer.

4. Analysis of Algorithms

A weighted graph would be displayed on the screen; the student would be required to point (perhaps with a mouse) to the edges that constitute a minimum spanning tree.

This type of question is particularly intriguing because there are two or three standard algorithms (that should be known to undergraduates) for finding such trees. The computer could easily detect which algorithm the student is using by examining the sequence of edges he selects. A number of outcomes are possible, including the following:

(i) The student fails to construct a tree at all. In this case he receives a zero score and is then asked very basic questions about trees and graphs.

(ii) He constructs a non-minimum tree in an arbitrary manner. In this case, the computer could suggest a sequence of modifications to the tree to convert it into a minimum one. In other words, the computer is giving him hints; when he chooses to, the student could ask that the hints be stopped and that he be allowed to finish the job himself. His grade would depend on the number of hints needed.

(iii) He is clearly using one of the standard, efficient algorithms, but made a minor mistake; he receives nearly full credit.

(iv) He gets the correct answer and receives full credit.

A Model for Assessing Undergraduate Learning in Mechanical Engineering

Jonathan Warren
Research in Higher Learning
Berkeley, California

Documenting What Students Learn in Undergraduate Mechanical Engineering Programs

Engineering education has a long history of self-examination. For the past century, major reports have been produced about every ten to fifteen years reflecting on the nature of the engineering curriculum--its purposes, content, organization, and length. Yet none of those reports described, or even asked, what students accomplished in completing an engineering program. The critical issue appeared to be what was offered, not what students learned.

What is taught, how intensively, for what length of time, in what way, using what resources--all these elements of an educational program unquestionably influence what is learned. They properly underlie judgments of educational quality by accrediting agencies and other educational authorities in this country and abroad. Yet because those examined elements say nothing about actual student accomplishments, they cannot be drawn on to resolve the issues in engineering education that have persisted for decades. Those issues include the relative emphasis to be given design problems rather than theoretical science, how much laboratory work is desirable, the importance of cooperative education, the impact of a large graduate program on undergraduate education, the importance of industrial experience in the faculty, the effects of the humanities and social science components of the curriculum, and the impact of instruction in management or economics.

The need for information on students' accomplishments is illustrated in a recent article by an engineering dean and faculty member decrying what they saw as an overemphasis on theory at the expense of design (Kerr & Pipes, 1987). They asserted that today's engineering graduates are less capable of solving the practical problems engineers face than were graduates of prior years. They may or may not be right. What evidence exists is skimpy, impressionistic, and anecdotal.

If they are right, and the proportion of the curriculum now committed to scientific theory should be reduced, which parts are most expendable? That question could be answered, but not well, by asking the faculty collectively to rank theoretical concepts or issues in importance. But learning is not compartmentalized. Understanding one theoretical concept or aspect of a problem facilitates the understanding of others. The less important concepts may provide links among more important ideas that would otherwise be less well understood. An optimal mix of theory and practice might be found if its boundaries were not too sharply specified. It will be recognized only in what students learn and in the gaps in learning that appear with different curricular structures.

Information on what engineering students in general learn is useful in limited ways but is too weak to support important decisions. The following apparently sensible question illustrates the difficulty in understanding information about general student achievement. "What capabilities have been acquired by graduates of engineering programs oriented toward design that have not been acquired by graduates of theory-oriented programs?" If the average "design" graduate is found to be more adept than the average "theory" graduate in solving the "real-world" problems faced by practicing engineers, that would seem to answer the question. But would the better and poorer students show similar differences? Toward either end of the distribution of achievement, the relative advantage of the "design" graduates might disappear or be reversed. Clearly, the distributions of the two groups in their ability to solve practical engineering problems would need to be examined.

Engineering students are neither a passive nor homogeneous group. They learn much that is relevant simply by observing their professors. They are also selective with respect to what is formally presented, learning best the material that meshes comfortably with what they already know and neglecting other material that seems unimportant or strange, or simply misses their attention. The nature of their selectivity varies, some students focusing intently on material that others barely notice because of differences in their prior experiences or personal inclinations. And some students are just more competent, more diligent, better organized, or less distracted. No one expects the capabilities acquired by the average graduates of a program to be the same as those acquired by the best students. But those differences are rarely described qualitatively. What have the better students accomplished that the ordinary students lack? Are those differences uniform? If not, how do the top students, or the ordinary students, or students who differ in learning style, or any number of other characteristics, vary in their engineering capabilities?

A report by the Center for Policy Alternatives (1975) wondered whether "...the 'quality' of a curriculum could be discussed only in the context of a given student or type of student" (p. 37). That view is too pessimistic. Some general statements can be made with confidence about all the graduates of any school of engineering. Most graduates successfully enter the job market as engineers, and many are successful in graduate engineering programs. Most programs can, with justification, be said to be effective. But how effective, and in what ways and for what students, and could some aspects of a program be readily improved?

The usual generalities about quality are not adequate for decisions about specific issues of curriculum and learning. Perhaps we can begin to replace many of these statements about engineering education with more clearly focused and documented assertions about substantive issues of student accomplishment. This report is addressed to that issue.

The argument is sometimes made that variations in learning associated with student differences are beyond the control of the faculty or institution; that once the curriculum has been organized and presented, the rest is up to the students. But some ways of structuring the material and some ways of presenting it are more effective for certain kinds of students and less so for others. To assert that engineering curricula should give more attention to solving practical problems and less to scientific theory, or the reverse, requires evidence of the effect of each curricular form on what various kinds of students in various kinds of circumstances learn. That kind of evidence is not now available.

Other disciplines have been no better than engineering in their inattention to educational accomplishment. A classic study of the early 1930's, which pointed out that the average sophomore students in some institutions were more knowledgeable than the average seniors in others, was a rare exception (Learned and Wood, 1938). That study has been greatly admired and widely quoted for half a century but has not been repeated, updated, or improved upon, even though procedures for assessing academic achievement have advanced appreciably. Thirty years later, a major book on the outcomes of undergraduate education was published with no mention of what students might have learned in college (Feldman and Newcomb, 1969).

More recently, a sweeping examination of undergraduate education, with chapters on fourteen fields of study including engineering, had virtually nothing to say about what students in those fields learned (Chickering and Associates, 1981). Personal, moral, ethical, vocational, and general intellectual development, even growth in the capacity for intimacy, were examined. But the primary educational purpose of faculty members, departments, and colleges--helping students understand some portion of a selected field of study and something of the broader context in which it is embedded--was not.

The personal and social development of college students is important; it deserves attention. The need of engineers for an education broader than knowledge of their technical field was asserted by Karl Compton more than fifty years ago (Compton, 1932) and has been repeatedly recognized by engineering educators ever since. But the heart of engineering education is the understanding of concepts in the physical sciences, their mathematical manipulation, and the ability to apply technical knowledge to the solution of engineering problems. Those broad phrases include a range of diverse and complex kinds of learning that engineering students reach to varying degrees.

A number of fundamental questions about engineering education have gone unasked. What theoretical concepts, mathematical tools, and problem-solving capabilities can most graduating engineering students be expected to have learned? What variations in that body of learning appear among students who follow different patterns of courses in their junior and senior years? How has that body of learning changed as the curriculum has accommodated new developments in science and technology? What have the best students learned that the ordinary students have not? Is there a general answer to that last question regardless of the pattern of courses taken? These neglected questions point toward a need to understand better what students in fact learn rather than being satisfied with what can be found in a program's list of courses and syllabi.

Purpose of the Project

The basic function of indicators of learning in the present context is to answer the question, "What does it mean to be educated at the baccalaureate level in a particular field?" A systematic description of the learning associated with the various combinations of courses students take seems as though it would answer that question, although its compilation might be complex and difficult.

The purpose of this project was to examine the feasibility of compiling information from faculty-developed course examinations into composite indicators of what students have learned in the various combinations of courses they have taken. To accomplish that purpose, information was gathered on

- The kinds of knowledge or capabilities faculty members in mechanical engineering want their undergraduate students to acquire;
- The patterns of courses that constitute several different mechanical engineering curricula;
- The nature of the course examinations commonly used; and
- Present indicators of engineering accomplishment in this country and abroad.

Despite the focus on mechanical engineering, the possibilities examined in the project are not limited to a particular field. The examination procedures in engineering differ in substance but not in their essential form or procedures from those in any field. What can be accomplished with the quantitative problems of engineering exams can probably be accomplished with the essay questions of other fields. Engineering was selected because it is one of the two largest professional fields in undergraduate education, and the professional fields are those in which issues of student achievement and educational quality may be clearest. Mechanical engineering was selected because it is the second largest engineering specialty in terms of enrollment and has a more standard and well-defined curriculum structure than electrical engineering, the largest of the engineering specialties.

Requirements for Indicators of Learning

At least two requirements must be satisfied by any practical indicator of learning. One involves the technical quality of the indicator measuring device; the other the time, cost, and complexity of its use. What does the indicator mean? Is it usable?

Substantive content

With most educational tests or other indicators of students' academic accomplishments, the question of whether the test accurately reflects the content of the relevant educational program is an issue of validity. But rarely does a test report describe the capabilities that led to whatever result was achieved.

Any indicator of learning that is widely used must provide information on the substance of what has been learned. Although that seems an obvious requirement, it is not true of the two most common current indicators--grades and standardized test scores. Both state students' relative levels of accomplishment in vaguely defined areas of learning--such as thermodynamics, or engineering in general--with respect to unknown groups of other students. The nature of the capabilities represented by a grade or test score is not known.

Scores and grades are sufficient for the limited purposes for which they are intended, which are to indicate educational preparation acceptable for more advanced study and acceptable completion of a segment of an educational program. Retrospective evaluations of the success of a program require information that describes the qualitative nature of what was learned and the ways that substantive learning differed with different kinds of students.

Course examinations, laboratory work, projects, reports, and other evidence of learning that faculty members routinely use do carry substantive meaning. What the students have and have not accomplished is apparent. When those observations are converted into grades, however, the substance is lost. Nothing can be said about what the students with better grades or higher test scores have learned that those with lower grades or test scores have not.

Gathering substantive information on what has been learned, despite its obvious value, presents at least one problem--how to describe the diversity of what students learn. Undergraduate learning varies widely in content and level of accomplishment. Students in the same department at the same institution do not necessarily graduate with the same kinds of understanding and capability. Among different institutions, the variability is still greater.

A recent report on engineering education grouped colleges of engineering into two "tiers" (National Research Council, 1986). The first tier consisted of colleges of engineering in universities with large graduate programs and extensive research support. The second tier included the other 90 percent or more of accredited engineering schools. The two tiers differ in their purposes, size, resources, and faculty and student characteristics. Engineering students in the two types of institutions encounter different kinds of programs and almost certainly differ in the qualitative nature of their accomplishments. Yet such differences, if they exist, are not shown by differences in their scores on a standardized test, and certainly not by their relative grade-point averages, which may be higher in less demanding programs.

A major difficulty with indicators of students' substantive accomplishments is in aggregating, summarizing, and comparing them. Narrative descriptions of what students have learned in a course are used in a few institutions, and summary statements are produced from them. But the summaries are difficult and time-consuming, and lose much of their importance when one person aggregates the descriptions written originally by a number of faculty members. If substantive indicators of learning are to be feasible, simple procedures must be available for cumulating, summarizing, and comparing the results.

Validity

Validity, as used in the assessment of educational accomplishment, is the strength of the evidence supporting the inferences drawn from the assessment. It is tied to the inferences, not the procedure itself. It may take different forms, with the same assessment procedure having different values depending on the nature of the desired inferences. For example, because the Graduate Record Examinations are intended to support inferences about the ability of applicants to succeed in graduate school, their validity can be indicated by the strength of the relationship between GRE scores and graduate school grades, completion of a graduate degree, or other indicators of graduate school success. Their validity may be quite different and would require different kinds of evidence if they were used to infer how well students had accomplished the purposes of an undergraduate engineering program, since those purposes are not limited to preparation for graduate school.

Scope and relevance. Indicators of learning are retrospective, describing what students have accomplished academically. Faculty members use them to evaluate student progress and guide whatever adjustments in instruction may be indicated. Department heads, deans, and vice presidents use them to evaluate the effectiveness of curricular structures and instruction and the adequacy of resources. For those purposes, the information provided must have both scope and relevance. It must indicate the full range of learning acquired by the students in completing a program--engineering capabilities as well as knowledge of theoretical science and mathematics. It must also be relevant to the particular areas of specialization that reflect the preferences and capabilities of the faculty and in which students focus their energies. Because widely used indicators must represent the material most commonly taught, broadly applicable standardized tests such as the GRE often miss the major strengths of some departments. Inferences about educational quality lack validity to the extent that the information on which they are based is incomplete or lacks relevance to the particular courses that make up students' programs.

Reference groups. The validity of inferences depends partly on their context. Inferences about quality are particularly dependent on the context from which they are drawn. Context is provided by reference groups, with a greater variety permitting more informed inferences. Whether the observed accomplishments of students completing a particular group of courses are remarkably good or fall far short of what might be expected can be understood only in relation to similar accomplishments of other students, the nature of the programs they went through, and the capabilities they had on entering the programs.

Large reference groups, sometimes national in scope, are often cited as one of the strengths of standardized tests. A national norm against which a department can compare its students' achievements is appealing. Yet the value of "national" norms is limited. The characteristics of the students and their institutions are unknown and so diverse that no real institution matches them. Comparisons based on them carry little useful meaning. More useful than large norm groups for inferring the quality of the accomplishments of any group of students are results from a few other groups having known characteristics. Ideally, some reference groups would be similar to the group being evaluated and others would differ from them in known ways.

Reliability

Any useful indicator of educational achievement must provide results that are consistent, that are not affected by extraneous circumstances or particular applications. The grades assigned to essay tests, for example, may fluctuate with the grader and the circumstances of the grading. Grades assigned to quantitative problems may be more consistent than essay grades but may still be questionable. Test grades are also inconsistent, or unreliable, if the questions or problems on which they are based are ambiguous or if students misinterpret them for any reason. Course grades, however, which represent a composite of several observations of performance, are usually reliable, as indicated by the consistency with which students tend to be A students in whatever course they are in, as do B and C students. If course exams are to be a reliable basis for broad indicators of learning, they will have to produce consistent results that are graded consistently.

Structure of the Project

The premise which the project was designed to examine is that without inordinate effort or cost, information from regular course examinations can be aggregated into broadly useful indicators of learning. For that to be possible, some assurance would be needed that faculty-developed exam problems in mechanical engineering allowed valid inferences about the substantive accomplishments of undergraduate mechanical engineering students.

Sources of data

The examples of courses and examination practices were drawn from four mechanical engineering departments in two State universities, one private university, and one private institute of technology. Each institution is in a different geographic region. One of the State universities is among the "first tier" universities--those with large graduate programs and substantial research budgets (National Research Council, 1986). The other State university, one of about 270 schools of engineering in the "second tier" group, was above the 75th percentile in undergraduate engineering student enrollment. The private university was slightly above the 25th percentile. The fourth department differed from the others in being part of an institute of technology, where the humanities and social science courses may be more closely related to the needs and interests of engineering students than they are in the other departments, and where the academic climate in the institution as a whole may be different. These four departments differed in size, resources, number of engineering courses offered, and research involvement of the faculty.

The characteristics of the four mechanical engineering curricula were identified from descriptions in the college catalogs, course descriptions of each engineering course prepared for the Accreditation Board for Engineering and Technology (the "ABET pages"), and transcripts of 50 recent graduates from each institution except the small private university (because its department was comparatively new, it provided transcripts of all its graduates, a total of 27). From the transcripts, the most common patterns of elective courses (technical and nontechnical) taken by mechanical engineering students at each institution were identified.

A total of 15 faculty members at three of the institutions were selected who collectively taught a broad mix of courses that represented the major course patterns identified in the transcripts. Each faculty member was interviewed by phone about what they wanted their students in that course to learn, and each provided a final exam from his or her selected course. The courses were all upper-division.

What students were expected to learn was thus examined in increasing detail, from general catalog requirements through course descriptions and syllabi to explicit statements by faculty members of what they wanted their students to accomplish in specific courses, and finally to the nature of the exams used to assess students' accomplishment of those desired kinds of learning. The diversity of the institutions and courses was such that an assessment procedure suitable to those varied settings would be broadly applicable.

Faculty Expectations for Student Learning

Four distinct kinds of learning desired by the faculty members interviewed appeared in their interviews and in what their exam problems asked students to do:

1. Understanding selected concepts of science and engineering--e.g., moment of inertia, blackbody radiation, viscous flow;

2. The ability to use the techniques or procedures necessary for manipulating or applying the concepts;

3. The ability to apply what has been learned to the solution of new problems in new contexts, or to combine concepts and techniques appropriately to solve practical problems of the "real world"; and

4. The ability to formulate an approach to loosely defined problems having incomplete information and no single answer.

Only the first two of these types of learning appeared with any frequency in the course descriptions, course goals, and syllabi in the ABET pages.

Understanding concepts

The most common inhabitants of course syllabi in mechanical engineering are the concepts to be learned and processes for using them. For example, the catalog description of a course titled "Heat and Mass Transfer" states that it covers "Steady and unsteady conduction, convection, and radiation; heat transfer; applications." Its goal is to "give students an ability to analyze various heat and mass transfer problems encountered in engineering as well as do preliminary design of heat exchangers." The topics listed in the syllabus include several concepts in addition to those mentioned in the course description and the use of "empirical equations for forced convection". Thus, a number of related concepts and their use in solving problems constitute the essence of the course.

Proficiency with techniques and procedures

The use of techniques and procedures includes mathematical procedures, which themselves often incorporate new concepts, such as the derivative of a function. In an engineering context, however, they are tools. Computer programming and use, as in simulation and computer-aided design; instrumentation and laboratory procedures; familiarity with standard tables; and other procedural skills are also among the techniques and procedures that constitute this kind of learning.

Ability to solve well-structured problems

The third objective--bringing concepts and techniques together in the solution of engineering problems--was described repeatedly by the faculty members as their primary course objective. It is a major requirement in design and is no doubt implicit in the "preliminary design of heat exchangers" of the heat and mass transfer course. The faculty members interviewed mentioned it when talking about the kind of learning that best discriminated A students from C students, or that best characterized their top students.

Since this ability was considered so important by the faculty members, three variations of it that appeared in the interviews are described below with illustrative quotations.

1. The ability to find links among concepts and between concepts and concrete situations.

"Strong students are able to take what they've learned in the course and connect it with other things they've learned in other courses. The less strong student would have learned how to do certain problems but wouldn't necessarily know how they relate to something going on in an oil refinery or a jet engine."

"The thing that makes the difference [between A and C students] is the ability to see the connections between the processes, the materials, and the design....If they understand the fundamental processes and can extend that to techniques necessary to produce something they haven't seen before, *that would characterize an A student.* The C student would be able to tell me a reasonable amount about the processes but would not be able to extend them to a new application or identify an appropriate process given some product."

2. The ability to apply procedures learned in one setting to new and unfamiliar situations.

"I also ask one question on every exam that they haven't seen before. They've all had the knowledge in the course to solve it....It's a natural extension of what they've had. They can conceptualize it and abstract it out....But C students don't make the jump from the concrete to an abstraction like that."

"In the exams some of the questions ask them to synthesize their knowledge. For example, I tell them, 'Here's a drawing of a part. You're the production engineer. Give me a proposal for manufacturing it.' It distinguishes between the people who have been repeating what they've heard me say and the students who really understand the process and can synthesize something they haven't seen before. It's important to address that level of learning where the person can apply what he's learned to a situation he wasn't familiar with."

3. The ability to formulate a problem as well as solve one that has already been set up.

"On an exam I flavor the problem so they have to discern what to do. They all know how to do something. If you set a problem up for them, the C students can do as well as the A students."

"The A students can take a word problem and formulate it; the C students almost never can."

"All the students are smart enough....The A students know where to focus their effort."

The following statement, in describing a distinction between one kind of accomplishment and another, includes elements of all three variations of this kind of learning.

> "Some students have a really good understanding of the concepts and mathematics. You write a complicated equation on the board and they appreciate it and can solve it. But then in the application part of the problem they're stuck."

Ability to solve poorly defined problems

The fourth kind of learning is an extension of the third but is listed separately because it was described by some faculty members as a prototype of the kind of problems faced by practicing engineers. It appeared in some of the exam problems submitted by the faculty members, but was most apparent in a few of the problems of the Principles and Practice of

Engineering Examination administered by the National Council of Engineering Examiners for the registration of professional engineers.

This kind of accomplishment, expected of advanced engineering students, requires a deeper understanding of the relevant concepts than is necessary in solving well-defined problems. It also requires imagination, flexibility, the ability to deal with uncertainty and to make informed assumptions. Cost and feasibility may impose constraints on the solution and affect judgments about its success. The best solution is determined not by whether it is correct but by whether it is optimal in some sense that may be chosen and defended by the student. Most of the problems students face in their courses differ, having a single answer that is either right or wrong.

Problems of this kind are considered representative of the design activities of engineers and are often called "design" problems. They appear in senior design courses but could well be introduced earlier in the curriculum.

Faculty Course Examinations

The mechanical engineering faculty members interviewed were selected because they taught a course that appeared in one of the common patterns of electives identified from transcripts. In addition to being interviewed, they submitted a course examination, stated what they wanted one or two of the problems to indicate about the students' accomplishment, and described typical student responses to those problems.

The problems and responses were discussed with the faculty members to determine whether the responses could be grouped into mutually exclusive categories that would reflect the most important kinds of student capability or understanding the faculty members wanted the problems to assess. For most problems, such categories were possible, although the requirement that they be mutually exclusive was sometimes difficult to meet. It often required merging some categories that were logically distinct but related. Since actual student answers were not available, the effect of combining categories could not be known.

Recording the substance of what has been learned

A common practice of faculty members in grading problems is to construct (explicitly or implicitly) a hierarchy of types of students' responses. This hierarchy may be based on awarding points for each correct element in the solution or on deducting points for each element of the problem that was missed. In both methods, the problem is organized into several distinct processes, each contributing to a score or grade. Different kinds and numbers of omissions or missteps result in lower scores.

Once a score is assigned, no record is kept of what the student had or had not done correctly in working the problem. If the processes identified for grading student answers were used as the basis for six to eight mutually exclusive categories, the distribution of students' solutions over those categories would identify the substantive nature of their successes and failures. Grading would consist of assigning a student's solution to the appropriate category, and the category descriptions would indicate what the students in each category had accomplished.

An example is provided by one of the problems for a course in compressible flow. It asked the students to calculate the mass flow rate of air from a reservoir through a converging-diverging nozzle, given the pressure and temperature in and outside the reservoir and the dimensions of the nozzle. A solution requires understanding the relationships among 15 or 20 concepts that can be organized into three sets--those related to (1) sonic and subsonic flow, Mach number, pressure, and temperature; (2) nozzle effects; and (3) mass flow, density, linear velocity, and their relationships to volume flow. It also requires knowing what relevant information is available in standard tables. Drawing on these sets of relationships and the given and tabulated data, the students first find the Mach number at the nozzle throat and then the mass flow rate.

Errors of understanding or execution can be made in working with each of the three sets of relationships, singly or in combination, and in using the standard tables. If the students' work with each set of relationships and the tables is judged successful or unsuccessful, all the possible combinations of successes and failures would require 16 categories. Some of those categories, though, will not be necessary because few if any answers that fit them will appear. Students who successfully negotiate the first set of concepts, for example, are unlikely to miss the third. Thus, four of the possible sixteen categories--those that include success with the first set of concepts and failure with the third--will not be used.

Categories can also be merged. Failures with the third set of categories and failure in using the tables correctly may both be rare. If the distinction between those two types of error were ignored, the sixteen potential categories would be immediately reduced to eight. Additional categories, though, may need to be added. Some important sources of error will be in the students' formulation of the problem.

The problem-solving ability that faculty members considered so important may be independent of specific areas of knowledge or understanding. Students who understand all three sets of relationships in the above problem and who can use the tables may still be unable to pull the various pieces together into a correct solution. An important additional category or two would indicate success or failure in merging the concepts and techniques of the problem into a successful approach to a solution.

In most problems such as that illustrated above, the important and informative ways students will go astray can usually be organized into about six or eight mutually exclusive categories.

For grading purposes, a grade can be associated with each category. The students receive the grade of the category to which their answers are assigned. Usually several of the partially correct categories will be assigned the same grade, with all six or eight categories spread over only three or four grades. In those circumstances, recording the distribution of answers over all the categories not only preserves the substance of the learning associated with the categories, but includes useful information about differences in what the students have learned that is lost when different kinds of partial success are awarded the same grade.

The argument is sometimes made that if instruction is successful, almost all the students will succeed. Almost all the answers to an exam problem will then fall into the single category of complete success. Although it is often realistic to assume that all the students in a course have successfully completed it, that does not imply that their accomplishments do not differ in important ways. Exam problems that fail to identify those differences do not provide enough information to justify the time students give working on the problem and faculty in grading it. More probing problems will identify the same kind of success, and areas of deficiency as well.

The result of a critical analysis of what students do with a problem will be a set of descriptive categories into which the students' responses can be quickly and accurately sorted. Time spent defining the most informative categories will be more than made up in the ease of grading by sorting the students' solutions into the categories.

Learning Attributable to Groups of Courses

As indicated earlier, a major premise of the project was that substantive records of student accomplishment drawn from course examinations can be cumulated into broader indicators of what students have learned. One of the appeals of grades is that they can be converted to numbers, cumulated, and averaged to produce a broad indicator of accomplishment. The price of that ease of cumulation is the loss of substance. A grade in an introductory course in fluid mechanics, for example, is indistinguishable from one in an advanced course in the design of thermal systems.

Recording exam results in the form of distributions of students across a set of substantive categories for each problem, with definitions of the categories, produces a descriptive record of the students' accomplishments. A major attribute of those descriptive categories, their ability to describe the diversity in what students learn, also makes their cumulation difficult. Most engineering students in the last two years of their programs take 50 or 60 examinations consisting of 300 to 400 problems. If 100 of those problems were graded by sorting the responses into 600 to 800 descriptive categories, the resulting statements of accomplishment would be too unwieldy to use easily. That problem can be reduced in two ways--by cumulating the results across patterns of courses rather than all possible combinations of courses, and through faculty members collaborating on exam problems to be used in more than one course.

Patterns of courses

The patterns of science and engineering elective courses taken by recent graduates in mechanical engineering from the four institutions described earlier were identified in their transcripts. The patterns consisted of clusters of science and engineering electives taken by approximately the same groups of students.

The course clusters were identified by calculating an index of similarity among all possible pairs of elective courses based on the proportion of students who took both members of a pair. The index ranged from zero, when no student took both members of a pair, to one, when every student who took either member of a pair also took the other. Cluster analysis (Sneath and Sokal, 1973) was then used with the matrix of similarity indexes to define clusters of courses that tended to be taken by the same students. In each department, a pattern of elective courses consisted of one or two clusters plus one to three additional courses not part of a cluster.

This procedure for forming clusters of courses and then grouping clusters and a few other courses into patterns is based entirely on students' choices. It may put courses together that have no inherent relationship with each other. The time a course is given or the popularity of the professor rather than the subject of the course may influence which students take it. Yet these clusters define similarities in students' programs and, presumably, in what they have learned. Identifying the accomplishments characteristic of students who have taken a particular pattern of courses--one or two clusters plus a few more electives in addition to the required courses--simplifies aggregating the exam results into broad indicators of what has been learned. A complete upper-division program in these four departments consisted of six courses required by all four departments, five to eight additional required courses that were not common to all four, and five to eight elective courses organized into clusters.

The clusters showed similarities and differences across the four departments. A cluster centered around thermal engineering appeared in three of the four departments. In the fourth department, the dominant cluster combined courses in propulsion and combustion engines, and a similar cluster appeared in another department. The clusters varied, though, from department to department. In one department, a course in compressible flow was clustered with three courses on aerodynamics and aerospace engineering. In another, compressible flow was combined with two courses on heat exchange processes and one on combustion. A question this illustrates is whether similar courses in compressible flow would have different effects depending on the nature of the related courses taken.

Cumulative effects of individual courses

The way the answers of any group of students to a given exam problem are distributed among a set of mutually exclusive categories describes in detail the collective accomplishments of that group in the areas the problem addresses. The group may be constituted in whatever way is most useful--all the students in a particular course, all those who have or have not had a year of co-op experience in industry, or all the students in several related courses. Recording examination results in terms of the distributions of students' answers to each problem permits subdividing or cumulating the results of instruction without losing their substantive meaning.

Accumulating exam results from several related courses may be illustrated with one of the clusters described above. In one of the departments, four upper-division elective courses often taken by the same students were Introduction to Aerospace Engineering (ME311), Theoretical Aerodynamics (ME320), Gas Dynamics (ME421), and Aerospace Design (ME452). ME311 and ME320 are taken in either the junior or senior year, the other two in the senior year. Although the junior-year courses tended to precede the other two, none is a prerequisite for any of the others, and both senior-year courses were occasionally taken without either of the earlier ones. Of the students who took any one of the four courses, more than 80 percent took at least one other, and 20 percent took all four. The four courses are clearly related and should be expected to build on each other.

The cumulative effects of the four courses should be apparent in the relevant understanding and capabilities of students who took different subsets of the four courses. Six problems might be devised collaboratively by the professors teaching the four courses that would assess areas of understanding common to more than one of them. The problems need not be identical. They might require different combinations of understanding or capability as long as some portion of their requirements and one or two of the categories for grading were common to the four courses. Two or three of the problems could be inserted into the final exams of all four courses one year and the others the next year.

If the results from all four courses were broken into groups of students according to which combination of the four courses each student had taken, comparisons of the percentages of each group in the common response categories would indicate the contribution to the students' accomplishments of each course individually and in combination with the others.

A more accurate indicator of possible interactive effects of two or more of the courses would be provided if the collaborating faculty members constructed common exam questions focused explicitly on kinds of learning they would expect to be enhanced by two or more of the courses in combination. The grading categories would also be defined to reflect those expectations. If any two or more courses do in fact interact in their effects, the students who had taken different combinations of courses would be expected to distribute themselves differently over the various types of responses to each of the common problems. Those differences will be observable even when the two subgroups of students have the same average grade on any of the problems or on the exam. Qualitative differences in what has been learned do not always result in different grades.

Construction of the categories to be used in grading the problems is clearly important to the value of the information the problem provides. They should be defined so the distinctions among the categories correspond to the most important distinctions among the students. That implies attention to both the kind of student accomplishment of most interest and to the most salient distinctions that appear in the students' answers.

A single problem common to the exams of two or more courses can provide useful information but is necessarily limited in its scope. An understanding of the relationships among a few concepts relevant to several courses, or facility in solving a particular type of problem, might be identified in a single problem common to more than one course. More comprehensive information would be provided if, over the course of a term, each of several related courses that gave two midterms and a final exam included one or two common problems on each test. The scope of those eight to twelve problems would be broad enough to give reasonably clear information about how the courses interact in their effects on student learning.

Understanding concepts, familiarity with techniques, and problem-solving abilities are all involved in the objectives of most mechanical engineering courses. All three kinds of learning overlap. Some of the more important concepts appear repeatedly, while more specialized concepts are found in a single course. Techniques or procedures also vary in their generality, and facility in solving problems of various kinds may be the most general of the three kinds of learning. Different exam problems requiring the same problem-solving ability can be devised for different courses. One response category associated with that ability can then be established for each problem (and separate from the content of each problem), permitting comparisons of that ability among students who had taken different patterns of courses.

The problem on compressible flow described earlier can be used again as an example. More than 15 concepts needed to solve the problem were grouped into three sets to define categories of partially successful answers. An additional category was suggested to represent an understanding of all the concepts necessary to formulate a successful approach to the problem. The categories associated with any of those three sets of concepts, or the ability to formulate the problem, could be used with different problems that involved one or two but not all three of the same sets of concepts. A number of different problems, in a variety of courses, could be devised that require an understanding of the relationships among subsonic, sonic, and supersonic flow, Mach number, enthalpy, and entropy, as represented on a Mollier diagram, and the ability to use that understanding to formulate a solution to the problem. Among the six or eight categories of attempted solutions for each problem, one or two could be devised that were common to all the problems.

Combining the results of categorically graded problems

When a single group of students takes an exam consisting of several problems, a common way to grade and record the results is to assign each problem a maximum value, grade each problem numerically, and add the values received for each problem by each student to produce numerical grades for the test as a whole. Those grades can then be averaged to produce a mean score that characterizes the performance of the group. Some statement of the distribution of grades may also be used to further describe the group.

For a single problem graded categorically, the distribution of student answers over six to eight categories gives a simple and more informative description of the students' accomplishment than a mean score. When a single group of students takes an exam on which more than one problem is graded categorically, combining the results into a general description of the performance of the group as a whole requires as many statements of the distributions of students, and descriptions of categories, as there are problems. The complexity can be reduced, though, if some of the same categories are used in more than one problem. Whether the students of interest are a portion of a larger group or a combination of several groups, their collective performance is indicated in the proportional distribution of their answers over the set of descriptive categories. Such a record is far more informative than the students' grade distribution and a course title.

Uses of cumulated distributions of categorical results

The use of common response categories with different problems makes possible a wide range of reference groups against which the results of any single group of students can be evaluated. How an understanding of compressible flow differs when learned in a basic thermodynamics course, a course in air conditioning and refrigeration, or one in aerodynamics could be identified with different problems in each course that had a few common categories. The faculty members teaching each course could better evaluate their own students' accomplishments by comparing the proportions of their answers in the common categories against those of students in the other courses.

From one perspective, indicators of student learning are most valuable when used by individual faculty members to adjust their instruction to the particular successes and difficulties of their students. They may move the course along more quickly, slow it down, shift some emphases, or plan a new approach for next year based on their own students' performances in relation to those of other students in similar and related courses. From an administrative perspective, indicators of learning are most useful when they apply to large groups of students, such as all graduating mechanical engineering seniors.

Within a single institution, for example, five or six problems with incomplete data and no predetermined answer might be devised for use in several design courses and several courses with only a minor design component, all taken during the first half of the senior year, when some of the students will not have taken as heavy a load of design work as others. One or two of the problems could be included in an exam of each of the courses. The five or six problems might collectively include two or three different sources of difficulty in formulating a solution, using two or three different content areas. All the response categories would not necessarily be the same, but one or two would represent facility in formulating problems with uncertain answers. The problems would be relevant to each course, and their results would be used in assessing student progress and assigning grades just as those of any other exam problem would be. But the results from different courses where students had equivalent levels of design work could be cumulated by calculating the proportions of the answers for the total group that fell in the common categories.

General questions such as the relative importance of theory or design, or the value of laboratory work, might best be answered with information from more than one institution. The collaboration of faculty members in devising categorically graded exam problems could be extended across a few neighboring institutions or across similar institutions that are geographically dispersed. The process can go more quickly in face-to-face meetings but can also be accomplished by telephone or over computer networks. Several of the comparisons of exam questions in the current project were carried out with electronic mail.

Benefits in the Cumulated Results of Categorically Graded Exam Problems

A comprehensive record of what students have learned that is built from portions of the 100 or more course examinations they take in completing an undergraduate program has several desirable qualities that most other indicators of learning lack.

First, the results of categorically graded exam problems carry descriptive, qualitative, or substantive meaning rather than only a relative order of merit. This is a requirement met by few alternative indicators of learning. A high score on the GRE Engineering Test, for example, carries no information as to what contributed to that score beyond a general understanding of some of the knowledge expected of engineers applying to graduate school. Yet the diversity in the understanding and capabilities engineering students have acquired in completing an undergraduate program cannot be described without substantive information about the nature of those capabilities. This characteristic is particularly important in any judgment about educational quality.

Second, the frequency of exams given periodically in each course makes them far more comprehensive in the material covered than most "comprehensive" exams given at the end of a program or even at intermediate points. Comprehensive exams of some form are the major alternative to indicators of learning derived from course exams. In addition, the results of course exams are directly related to the purposes and content of the courses students have taken. Students in the same mechanical engineering department follow different patterns of elective courses, acquiring different kinds of specialized capabilities. Departments also differ in their areas of strength and emphasis even though the content of the first two years is quite standard. Upper-division courses reflect the particular interests and capabilities of a department's faculty. A common test used in all mechanical engineering departments, or even one designed for a single department, would miss those specialized areas of learning. Both comprehensiveness and relevance to particular departments and programs contribute to the validity of the inferences drawn from the results of the process.

Third, because the tests are integral parts of students' courses and the basis for grades, students typically give them their best effort. End-of-year tests that do not affect the students' grades, which are often slighted by the students, may give an unrealistically low indication of the students' collective accomplishments. On the other hand, if end-of-year tests are important to the students' academic records, many students will feel greater pressure than they would in routine course exams. The effects of that pressure are uncertain. It may boost the performance of some students and inhibit that of others. Again, the validity of the results is likely to be greater for the important and familiar course exams.

Fourth, the reference groups provided with course exams are potentially superior to those of alternative indicators such as standardized tests because they can be tailored to the characteristics of the students being assessed.

Fifth, reliability is a quality in which standardized tests are inherently superior to course examinations because they usually include many more questions, or items. That superior reliability, however, comes at the cost of validity because it is achieved by creating the test out of mutually related items. Such a test provides a reliable but very general score, which reflects only what is common to that large number of items. The diversity that is a dominant characteristic of American higher education is lost in a test that assumes complete homogeneity in curricula, faculty, and students. Problems on course exams cannot match standardized tests in reliability, but collaborative efforts among faculty members can bring them to a very acceptable level.

An important benefit (though not a requirement) of indicators of learning based on course exams is the opportunity they permit for providing feedback to students and faculty in time for corrective action to be taken. This presumably happens routinely with mid-term exams, and an effort to cumulate their results into a general indicator of learning would probably improve that process. Collaboration among faculty members in devising the problems and their grading categories would focus them on important aspects of learning while providing reference groups that would improve the understanding of the results. Few alternative indicators of learning permit timely feedback.

Finally, the cost of a general indicator of learning built from the results of course exams is quite small. Some effort is required by the faculty in the collaboration required to give their exam problems the qualities that will permit aggregation into general indicators of learning. The extent of that effort beyond what faculty already commit to course examinations however, is minimal. A major advantage in the course-based indicators is the lack of testing time required beyond what is already part of the typical academic program. Minor additional costs are incurred in the administrative process of cumulating the results of exam problems from a number of courses and in the collaboration among faculty members. The collaboration, however, has its own justification in the clarification of instructional purposes among the faculty members involved.

Alternative Indicators of Engineering Proficiency

Two tests are widely used in the United States to assess the achievement of engineering students, and a third assesses the competence of practicing engineers. None of the three was designed to assess the quality of an educational program in engineering. For that purpose they have serious limitations, although for their intended purposes they are probably effective. Since they are well-constructed tests developed by competent professionals, they merit examination as models for the assessment of achievement in engineering.

The three tests are the GRE Engineering Test, administered by the Graduate Record Examinations Board; and the Fundamentals of Engineering Examination, and Principles and Practice of Engineering Examination, both administered by the National Council of Engineering Examiners. The first two tests consist entirely of multiple-choice questions and are intended to be taken by engineering students or "engineers in training". The Fundamentals of Engineering Examination is sometimes referred to as the "Engineers-In-Training" or EIT exam. The Principles and Practice examination consists entirely of engineering problems. All three are designed to be used in every engineering specialty, but the Principles and Practice exam has separate sections for each of the four largest engineering specialties--chemical, civil, electrical, and mechanical engineering.

The capabilities required for successful responses to the questions and problems of the three engineering tests were examined in relation to the four kinds of learning faculty members described as major objectives of their courses--understanding concepts, knowledge of techniques and procedures, ability to solve well-defined problems, and ability to solve poorly-defined problems. Since the purposes for which the tests were designed differ from those instructional purposes, such an examination is not an evaluation of the tests. It does, however, suggest possibilities and potential problems in comprehensive examinations as indicators of learning.

Each multiple-choice question and quantitative problem in the most recent publicly available form of each exam was classified into one of four types corresponding roughly to the four types of learning. The four types, described below, differ generally in their conceptual complexity, although a wide range of complexity can be accommodated by each type.

Type 1. Requires the examinees to know the definition of a concept or identify an example of it;

Type 2. Requires the examinees to know and apply a simple principle, process, or relationship;

Type 3. Requires the examinees to understand the interaction of two or more principles or processes and apply them jointly; and

Type 4. Requires the examinees to make assumptions, select one from among several possible approaches to a problem, and justify their answer as preferable to others.

88

The first two types and occasionally the third are found in the multiple-choice tests. The third and fourth types appear as problems for which the examinee must construct an answer.

The GRE Engineering Test

A form of the GRE Engineering Test used in 1983 was examined. The purposes of the GRE are to help graduate departments of engineering evaluate their applicants and to help students assess their own preparation for graduate school. The engineering seniors who take it have their results sent to the graduate departments to which they have applied. The only way an institution can use the GRE to evaluate its own accomplishment is to ask its students to take the test and have their scores sent to it.

The GRE is a 2-hour-and-50-minute test with about 140 multiple-choice questions. About 90 of them, which produce a subscore in engineering, sample basic knowledge and understanding likely to be common to engineering departments across the country. It is based almost entirely on lower-division content, since the varieties of specialized study found in upper-division courses cannot be adequately represented in a single test and are not uniform across different schools of engineering. The other 50 questions produce a mathematics subscore and require knowledge of calculus, linear algebra, numerical analysis, probability, and statistics. All 140 questions produce a total score. Major characteristics of the GRE test are broad scope, lower-division content, and multiple-choice format.

The difficulty of the questions is reported as the proportion of persons taking the test who got each question right. The usefulness of that definition of difficulty depends on the preparation and purposes of the persons taking the test. Its value to any institution depends on the match between its students and curriculum and those of the total group of persons who took the test. The primary goal in constructing the final form of the test from a large number of possible test questions is to maximize the discriminating power of the test by providing a range of difficulty among the questions.

For the engineering subtest examined, a question answered correctly by 46 percent of the test takers was at the median level of difficulty. The most difficult 25 percent of the questions were answered correctly by 12 to 30 percent of the test takers; the easiest 25 percent, by 62 to 88 percent. The mean score on the engineering subtest was reached with about 40 percent of the questions right; on the math subtest with about 55 percent right.

About half of the questions on the engineering subtest were Type 2 questions, requiring the application of a principle or concept. About a quarter were Type 1 questions, requiring only the definition of a concept, and another quarter were Type 3 questions, requiring an understanding of the interaction of two or more principles to identify the correct answer to a problem. No Type 4 questions were included.

The Type 3 questions might be expected to be more difficult than those that simply require knowledge of the definition of a term. That was not the case. The three types of questions on the engineering subtest did not differ in mean level of difficulty, and each type encompassed the full range of difficulty.

Difficulty in the GRE test questions depended more on the familiarity of the content than on the complexity of the process required to reach a correct answer. When the questions were grouped by content, most groups showed mean difficulty levels between 40 and 54 percent correct. Questions in mechanics were at the upper end of that range, somewhat easier than those in thermodynamics, which were at the lower end. The most difficult of the content groups were the six questions related to properties of materials. Five of the six were answered correctly by 30 percent or fewer of the test takers in 1983. Four were Type 1 and the other two Type 2.

The difficulty of the questions on properties of materials suggests the importance of the period of time between studying a topic and taking a test on it. All six questions involved material usually studied in a sophomore course on properties of materials. The only easy question, answered correctly by 76 percent of the test takers, was the one in which the content was likely to have been encountered after completion of that sophomore course. This example suggests that some students might score higher on the GRE Engineering Tests at the end of their sophomore year than as seniors or graduates, since more of the material will be comparatively fresh.

The registration of professional engineers

The other two tests--Fundamentals of Engineering and Principles and Practice of Engineering--were both developed by the National Council of Engineering Examiners (NCEE).

They are used by State boards of examiners for the registration or licensing of professional engineers. The purpose of both is to demonstrate minimum competence and understanding necessary for the solution of engineering problems. That purpose is far different from evaluating the quality of an educational program, which cannot be accomplished with a test designed to distinguish between the incompetent and the minimally competent. Completion of the Fundamentals Exam, which may be taken as an undergraduate, is a prerequisite to taking the Principles and Practice Exam. In addition, state boards usually require about six years of engineering experience, of which four can be met by a bachelor's degree in engineering, before the Principles and Practice Exam may be taken.

The results of the NCEE tests are not available directly to schools of engineering. They may occasionally be used indirectly as a measure of the quality of a college of engineering when the institution examines the proportion of its graduates who qualify for registration. Yet the meaning of such a measure as an indicator of educational quality is uncertain for several reasons. The proportion of graduates from a particular institution who apply for registration and take the NCEE exams varies widely, and their characteristics are unknown. The content of both exams is limited to major themes in each of the four engineering fields tested, neglecting much of the specialization in upper-division engineering programs.

Another unknown element is the way each State's board of examiners uses the test results in conjunction with other criteria for registration or licensing. In most states, most engineering practice is carried out in companies that employ engineers and is therefore exempt from the State registration or licensing laws. And many employers of engineers do not require registration (Professional Engineering and Research Consultants, 1981).

The Fundamentals of Engineering Examination

The Fundamentals of Engineering Examination consists of 210 multiple-choice questions, 140 in a 4-hour morning session and 70 in a 4-hour afternoon session. It is intended to assess understanding of the basic and engineering sciences. A large proportion of its content is drawn from lower-division science, engineering, and mathematics courses, but some of its material is typically found in upper-division courses in fluid mechanics, thermodynamics, and electrical circuits. The 70 afternoon questions include 50 answered by every applicant and 20 chosen by the test taker in two of five 10-item sets, each in a different engineering specialty.

When the questions are classified into the three types for which multiple-choice questions can be used, their distribution is quite similar to that of the GRE. About one-third are Type 1; not quite half are Type 2; about one-fourth are Type 3.

No information on the comparative difficulty of the Fundamentals test and the GRE is available. Some of the Type 3 questions on the Fundamentals exam, particularly in the afternoon session, are more complex and probably more difficult than any of the GRE questions, but the time limitations on the Fundamentals test are more generous.

The Principles and Practice of Engineering Examination

The Principles and Practice of Engineering Examination differs from both the GRE and the Fundamentals exam. In each of the two 4-hour sessions, one in the morning and one in the afternoon, the examinee selects four problems from among 10 in the chosen engineering specialty. Most of the 20 mechanical engineering problems in the test form examined were in mechanical design, energy systems, and thermal and fluid processes, with one problem each in control systems, economics, and engineering management.

All the problems require several steps between the given information and the required answers. In most of the problems, the steps are not immediately obvious, which means the examinees must have a comprehensive understanding of the type of situation given to select the appropriate intermediate steps. In three or four of the problems, the path from the given information to the specified solution is comparatively short and direct, much like problems on exams in college courses. Yet this test goes far beyond the other two in the depth of understanding required.

About half the problems are Type 3, requiring specific steps to specific answers. The other half are Type 4, permitting alternative approaches and sometimes different answers depending on the approach taken. In a few, the approach taken must be explained and defended. In at least one problem, more information is given than can be used. Occasionally a judgment must be made and defended as to whether the solution, to which the selected approach led, is feasible.

This exam is the only one of the three discussed that included Type 4 problems, which assess the ability to formulate an approach to a loosely defined problem for which more than one solution is possible. Yet the requirement to solve only eight of twenty problems means that all the Type 4 problems may be omitted.

A Model of an Indicator of Learning

As has been repeated in several different contexts throughout this report, the proposed model of an indicator of what mechanical engineering students have learned can be built from individual problems on regular course examinations. The necessary modifications of the usual processes faculty members use in devising an exam and grading its results include, first, collaboration among several faculty members teaching different but related courses to devise problems that can be used jointly, and second, grading the results by sorting them into six to eight mutually exclusive categories that best represent the distinctions among the students' responses that are most relevant to the instructional purposes of the course. The collective performance of any group of students is described in the distribution of the group's responses over the set of categories and the descriptions of the category definitions--that is, the criteria that cause responses to be placed in each category.

The groups of faculty members collaborating on jointly used exam problems can be formed simply by identifying professors who teach courses that have some area of overlap either in their content or in the kinds of problem-solving capabilities or other academic behavior they are intended to teach. More useful, though, would be to identify clusters of two to four courses that the same groups of students tend to take. Those clusters of courses will be the ones that most often interact with each other in their effects on what students learn, either in a simply additive way or synergistically.

In any mechanical engineering department, from 4 to 8 patterns of courses are likely to appear in the ways students put together 20 or so upper-division courses to constitute the most highly specialized part of their 4-year programs. Those patterns will consist of combinations of 1 or 2 clusters of elective courses, 8 or 10 required courses, and 2 to 4 individual elective courses not part of a cluster. That manageable number of patterns, with a manageable number of components, will permit exam results from individual courses to be cumulated to describe, separately for each of the most commonly taken patterns of courses, the cumulative effects of an entire upper-division program.

The collaboratively developed problems would be included as portions of the regular examinations, one-fifth to one-third of either midterm or final exams. An occasional problem could be introduced alone, at any time, as a 15-minute exercise. When a problem is used in different courses, the understanding of what the different types of responses imply about students' capabilities is far greater than when the same problem is used in a single course.

Within a cluster of 3 or 4 courses, 12 or 15 problems could be scattered among the midterms, final exams, or occasional exercises of all those courses, assessing specific kinds of understanding or capability that two or more of the courses share, or assessing general kinds of problem-solving or other intellectual capabilities that are relevant to those and other engineering courses. Some of the results of those problems--the proportional distributions of students' responses--will be combined for different subsets of the group of three or four, and some will be combined for the total cluster. If each problem has 6 qualitatively different types of response, with several common to different problems, 50 or 60 different narrative statements, varying from quite specific to very general, could be made about the accomplishments of the students who took that cluster of courses. Each statement would be in the form of the proportions of students in the group who demonstrated each of the several kinds of understanding or ability, or their lack, that are represented by the categories.

For the individual courses, grades could be associated with each category. Given the category definitions, their associated grades, and where their own response fell, students would know immediately what they had done wrong and what they might have done to produce a better answer. While assigning grades is a necessary part of the examination process, and too often its only purpose, the definitions of the categories and the distributions of students across them carry more understandable and more complete information than do grades.

Some categories of answers will indicate general kinds of ability, such as the ability to identify the critical aspect of a problem involving an exchange of energy when all the relevant information is not known. Results from those problems will be capable of being cumulated across exam results from more than one cluster of courses. Over a period of 2 or 3 years, a body of information about results from courses, clusters of courses, and patterns of upper-division programs could be cumulated into several hundred descriptive statements about what different groups of students, following different curricular patterns and having different kinds and degrees of success, had learned.

Relation of the model to requirements for indicators of learning

In an earlier section of this report, a number of qualities were listed that should be satisfied if an indicator of learning were to be successful in answering questions about the quality of an educational program. The first listed, because it is so rarely achieved, is that the indicator leave a record of the substance of what has been learned, and that the record apply not to the total body of students as an undifferentiated group but to different kinds of students who go through the same program with different kinds of success. The categorical grading of exam problems accomplishes that without losing the ease of aggregating and disaggregating results for whatever groups of students are of interest.

The validity of an indicator is an important and complex issue that is a quality not so much of the indicator as of the inferences it permits. To be useful, inferences about what students have learned should refer to knowledge, understanding, and capabilities that were part of the instruction the students received. Evidence that a group of students does not understand some aspect of turbulent flow has little utility for indicating educational quality if that aspect of turbulent flow was not part of their instruction. Basing an indicator of learning on the material of a course important enough to be part of the exams of that course assures the relevance of the indicator.

Comprehensiveness is related to relevance. The exams of a typical course occupy 4 or 5 hours, adding to about 20 hours of testing time in the course of a semester, or 80 hours in the junior and senior years of an engineering program. If a quarter of each exam is given to shared problems graded categorically and cumulated, the results of those 20 hours of testing will be more comprehensive (and therefore permit more complete inferences about the learning that has occurred) than most comprehensive exams.

One common source of evidence of validity in educational tests is expert judgment. In studies of test validity, faculty members who know the material are asked to judge the importance and relevance of each item of knowledge or understanding tested and the accuracy with which a particular test question represents it. When faculty members collaborate in devising a mutually useful exam problem, they necessarily conduct that kind of validity analysis, challenging and defending the usefulness or meaning of each problem, and the students' probable responses to it, with one or several other experts.

Finally, the reference groups are both relevant and known. The collaborating faculty members work with each other because they know what the other is teaching and the kind of students they have. And they have the added benefit of second guessing, which is not possible with norm groups for standardized tests. They can discuss with each other, after the results are in, why one group of students should have had so many responses of a particular type while another group did not.

Reliability is also strengthened by collaborative efforts. There is no reason to think the reliability of a composite indicator built from a number of statements of student results is any lower than the reliability of grade-point averages, built from similar kinds of faculty-constructed tests.

Two fundamental questions were stated at the beginning of this report that need to be answered about any indicator of learning. The first, "What does it mean?", is answered by evidence of validity. The second, "Is it usable?", has been answered in part in terms of time and cost, both small if only the actual administration of the test is considered. But collaboration among faculty members takes time, and so does devising and verifying the categories to be used in grading and the grading itself.

Two people collaborating to devise an exam problem they both can use will probably spend more time than either would alone devising his or her own problem. The added time should not be great, though, and as it increases so should the quality of the exam problem produced. Information is not available on how much time faculty members typically spend devising exam problems and how much more might be needed if they did it collaboratively.

The time spent reviewing student answers to see whether the anticipated categories of response really work is time not needed for the usual kinds of exams. Experience with essay questions as well as quantitative problems suggests that reviewing 20 to 30 answers will turn up all the types of responses to a question or problem that will appear frequently enough to be useful. It will take 1 to 2 hours, with the process accelerating as it goes. Once the categories have been identified, revised, clarified, and unambiguously defined, further answers are graded, or sorted into the categories, at an average rate greater than one per minute. One-half to two-thirds of the answers from any single class will be quite representative of the three or four most common types of response and will be assigned to the proper category after little more than a glance. Others will require more careful reading. A few, perhaps 1 to 3 percent, will not fit any of the categories no matter how carefully constructed. They should not be forced but should be graded individually and omitted from the collective record of the group as a whole.

The categorical grading of examination questions is not a new procedure. Norman Frederiksen has been working with the process at Educational Testing Service since the 1950's (Frederiksen, 1986). A project, involving representatives of 15 California colleges and universities, in devising both essay questions and quantitative problems was carried out by this author more than 10 years ago with results reported in an unpublished paper (Warren, 1978). The National Assessment of Educational Progress has been reporting its results in terms of distributions of responses to individual test questions for about 20 years (Messick, Beaton, & Lord, 1983). More recently, Lars Dahlgren has reported a study of the categorical grading of responses to oral questions, describing how "the prevalent qualitative variation can be described through a number of content categories [with] group differences...described through differences in distribution over the different categories of answer" (Dahlgren, 1984, p. 64).

Documenting the substance of what students have learned seems so indispensable to any analysis of educational quality that statements of what students have and have not accomplished (derived from their proportional distribution over categories of responses to exam questions) should have great appeal.

References

Center for Policy Alternatives. (1975). Future Directions for Engineering Education. Washington, DC: American Society for Engineering Education.

Chickering , A. W., and Associates. (1981). The American College. San Francisco: Jossey-Bass.

Compton, K. T. (1932). "Cultural aspects of engineering education." The Journal of Engineering Education, 23, 69-76.

Dahlgren, L. O. (1984). "University studies and concept of reality--A study of effects of education." Western European Education, 16, 61-70.

European Federation of National Associations of Engineers. (1975). Standards for Engineering Qualifications: A Comparative Study in Eighteen European Countries. Paris: Unesco Press.

Feldman, K. A., and Newcomb, T. M. (1969). The Impact of College on Students. Vol. I. An Analysis of Four Decades of Research. San Francisco: Jossey-Bass.

Frederiksen, N. (1986). "Toward a broader conception of human intelligence." American Psychologist, 41, 445-452.

Herbst, L. J. (1987). Personal communication, Dec. 30.

The Institute of Electrical Engineers. (1985). Guidelines for Engineering Applications EA1 and EA2. London: The Institute.

Japanese Colleges and Universities 1987. (1987). Tokyo: Maruzen.

Kalela, A. (1986). "Progress in co-operation for the implementation of the European Convention on the Recognition of Studies, Diplomas and Degrees concerning higher education." Higher Education in Europe, 11, 6-10.

Kerr, A. D., and Pipes, R. B. (1987). "Why we need hands-on engineering education." Technology Review, 90, 34-42.

Learned, W. S., and Wood, B. D. (1938). The Student and His Knowledge. New York: The Carnegie Foundation for the Advancement of Teaching.

Messick, S., Beaton, A., and Lord, F. (1983). National Assessment of Educational Progress Reconsidered: A New Design for a New Era. Princeton, NJ: Educational Testing Service.

Nast, M. (1986). "On the problems of comparison of curricula, examinations, graduation from higher education and academic programmes and degrees." Higher Education in Europe, 11, 13-18.

National Research Council. (1986). Engineering Undergraduate Education. Washington, D.C.: National Academy Press.

Perry, P. (1987). "Accountability and inspection in higher education." Higher Education Quarterly, 41, 344-353.

Pratt, J. (1987). Editorial: "Accreditation and innovation." Higher Education Review, 20, 3-6.

Professional Engineering and Research Consultants. (1981). Engineering Education and Licensing in California: A Report to the California Postsecondary Education Commission. Commission Report No. 81-19. Sacramento: California Postsecondary Education Commission.

Sneath, P. H. A., and Sokal, R. R. (1973). Numerical Taxonomy: The Principles and Practice of Numerical Classification. San Francisco: Freeman.

Statistical Yearbook 1987. (1987). Paris: Unesco.

Warren, J. R. (1978). The Measurement of Academic Competence. Berkeley, CA: Educational Testing Service.

Model Indicators of Student Learning in Undergraduate Biology

Gary W. Peterson and Patricia C. Hayward
The Florida State University

According to Bauer (1966) indicators are "statistics, statistical series, and all other forms of evidence that enable us to assess where we stand and where we are going with respect to our values and goals, and to evaluate specific programs and determine their impact" (p.1). The purpose of using indicators of program outcomes (in this case student learning) is to provide information for planning future policies, taking into account a wide range of important intended and unintended consequences (Scriven, 1972). At present, the major indicators of student learning in undergraduate biology consist primarily of a listing of the courses taken, grades earned in those courses, and scores earned on standardized tests such as the Graduate Record Examination Biology Test or the Medical College Admissions Test (MCAT).

Although such indicators have utility for admission to graduate or professional schools, they probably have little merit as measures with which to evaluate the effectiveness of undergraduate programs with respect to student learning in undergraduate biology. Generally, they reflect neither the comprehensiveness, nor uniqueness of individual programs and their curricula. If indicators that measure such attributes could be identified or developed, they could be useful for clarifying the objectives of individual programs, facilitating decisions pertaining to curricular development within programs, permitting greater accountability for resource allocation, assisting in the recruitment and selection of students, and helping high school students identify colleges offering programs compatible with their interests and abilities.

The scope of our investigation was limited to the identification of indicators of student learning in the undergraduate major of biology. We were not interested in recapitulating prior efforts in deriving indicators of departmental quality (Hagstrom, 1971; Young, Blackburn, and Conrad, 1987), institutional quality (Brown, 1970; Miller, 1979; Oakes, 1987) or graduate school prestige (Jones, Lindzey, and Coggeshall, 1982; Conrad and Blackburn, 1985). Nor were we interested in deriving indicators of general learned abilities (Alverno College, 1976; Peterson and Watkins, 1979) or in the long-term effects of student learning such as employment rates, salaries, or health (Jones, 1985; Soloman and Taubman, 1973).

The derivation of a set of indicators to assess student learning in biology requires enumeration of both (a) student learning outcomes and (b) ways learning outcomes can be measured. The form and substance of the ways these two aspects are integrated are referred to as a "model" (Kaplan, 1963; Land, 1975). Student-learning outcomes on can be analyzed in terms of four principal components: knowledge; cognitive skills; technical skills; and affective learning. Types of assessment, on the other hand, can be analyzed in terms of a continuum from proximal (i.e., direct measures) to distal (i.e., highly inferential measures) (Landy and Farr, 1983). The "model" we have developed consists of a two-dimensional matrix with student-learning components on one axis and proximal to distal measures (indicators) on the other.

Indicators can be identified or developed for each of the cells of the matrix. Once the generation of indicators is completed, they can be aggregated to produce comprehensive tests, which are direct measures of desired student learning, or faculty and student surveys, which assess student learning indirectly. The purpose of this paper is to show how indicators of learning can be developed that reflect the broad range of important learning outcomes in biology and to present several prototypical measures that might be used to assess the extent to which students are mastering these outcomes.

Identification of Model Indicators

A five-stage paradigm, designed to structure the process for developing model indicators of student learning, was adapted from an earlier work on the development of the Academic Program Evaluation Paradigm (APEP) sponsored by the American Association of State Colleges and Universities (Peterson, 1982). The five stages, which may be appropriate for the identification of indicators in any academic discipline, are as follows:

> **Stage 1: Define learning outcomes.** Concise statements are formulated to describe the broad range of knowledge components, cognitive skills, technical skills, and attitudes and values that should be learned by the best graduates of a program. Such statements, which reflect the results of teaching, are referred to herein as "learning outcomes." Statements of learning outcomes should be both discrete and comparable in level of abstraction.
>
> **Stage 2: Analyze organizational variants.** The purpose of this phase is to ascertain the relationships between types of programs and the importance of various learning outcomes. The analysis may consist of a series of contrasts between public and private institutions, large and small programs, Ph.D.-granting and non-Ph.D.-granting institutions, and top-rated and "normal" programs.
>
> **Stage 3: Determine generalizable and specific outcomes.** Core outcomes are the cognitive skills, laboratory skills, and values that are important to all types of programs, whereas specific outcomes are those that are unique in their importance to certain kinds of institutions or programs.
>
> **Stage 4: Identify model indicators.** This stage encompasses two phases: (1) an evaluation of existing measures of cognitive skills, laboratory skills, and values and (2) the development of model indicators. The first phase entails the analysis of existing tests and measures, and the second phase involves the identification and/or generation of assessment methods that will yield indicators for each cell of the matrix of learning outcomes.
>
> **Stage 5: Design specifications for model indicators.** In this last stage, student learning indicators produced in Stage 4 are used to develop tests and surveys that can be used to gather information pertaining to the extent to which students are achieving the important learning outcomes.

The remainder of this paper describes how the five-stage process paradigm was implemented from a national perspective in biology and presents several prototype measures of student learning developed from the process.

Method of Implementation

Stage 1: Define learning outcomes. An interinstitutional group of 10 faculty members from five colleges and universities in Florida assembled for a 2-day forum to describe the cognitive skills, laboratory skills, uses of equipment, and values and attitudes that should be mastered by their best graduates with majors in biology. The biology programs represented were from Florida State University, a public, Ph.D.-granting program; Florida International University, a public, non-Ph.D.-granting program with significant minority enrollment; Florida Atlantic University, a public, non-Ph.D.-granting program; Miami University, a private, Ph.D.-granting program; and Stetson University, a private, non-Ph.D.-granting program. The two faculty members from each program included the department chair and one faculty colleague with an active research program. The 10 participants were subdivided into three workgroups and given the directive, "At graduation, what do you think your best graduates should know, think, do, believe, or value?"

A total of 133 statements of knowledge components, cognitive skills, technical skills, equipment skills, values, attitudes, questions, and issues were generated. These statements were aggregated into categories and subcategories. Redundant statements and statements subordinate to higher-order statements in the respective categories were eliminated. Six statements were subsequently added to the pool from lists of outcomes contained in "Evaluation of Learning in Science" (Klopfer, 1971) or "Criteria for Excellence in Biology" (NSTA Report, 1987). The statements were edited, then reviewed by members of the interinstitutional workgroup. The final list of outcomes contained 34 cognitive skills, 53 laboratory skills or uses of equipment, and 13 values and attitudes (see table 1).

Stage 2: Analyze organizational variants. A survey containing the learning outcomes formulated in Stage 1 was sent to the department chairs of a national random sample of 192 programs drawn from a population of 1,360 colleges and universities. The list of institutions was obtained from Peterson's Annual Guides/Undergraduate Study: Guide to Four-Year Colleges (Lehman, 1987). The population excluded Bible colleges the theological seminaries; schools specializing in accounting, art, medical technology, music, pharmacy, engineering, and ethnic studies; and colleges with enrollments of less than 100. Of the 192 biology programs, 6 spanned more than one department. In these cases, the chair of each department within the program received a survey. In all, 199 department chairs received questionnaires. The survey, The National Inventory of Student Learning in Undergraduate Biology (Peterson & Hayward, 1987a), asked the chair to rate each learning outcome in terms of its importance and whether it was taught in the core curriculum (yes, no) or elective curriculum (yes, no).

On the advice of the interinstitutional workgroup members, the same survey was sent to 66 department chairs of the programs rated in the top 50 programs by the Gourman Report (1985). The rationale for the inclusion of this set of programs for study was that the identification of core learning outcomes should be based not only on a normative consensus, but also on the values possessed by the most highly regarded programs.

Table 1.—List of cognitive skills, values and attitudes, and laboratory skills and techniques

Cognitive Skills

Basic Knowledge

1. Specific facts measured by standardized tests
2. Concepts measured by standardized tests
3. Relationships between concepts
4. Principles of classification
5. Chemical bases of biological phenomena
6. Physical laws related to biological phenomena
7. Principles of behavioral and social science related to biological phenomena

Observation and Measurement

8. Describe observations appropriately
9. Select appropriate measuring instrument
10. Use metric system
11. Use scientific or statistical conversion tables
12. Determine measurement error

Inquiry and Problem Solving

13. Demonstrate knowledge of scientific trends
14. Apply moral/ethical issues to scientific trends
15. Conduct a literature review
16. Propose a research study
17. Formulate a hypothesis
18. Design a controlled experiment
19. Use population sampling in designing experiment
20. Conduct lab or field experiment
21. Design a data collection sheet
22. Construct graph, table, or chart to express relationships among variables
23. Express relationships among variables in the form of math equations
24. Use statistics to demonstrate relationships between and among variables
25. Draw inferences from a data set
26. Formulate appropriate generalizations
27. Write a scientific report
28. Give an oral report
29. Relate findings to scientific theory

Critical Thought

30. Describe arguments for opposing sides of a controversy and defend a position
31. Detect a false claim
32. Use different modes of thought
33. Differentiate fact, opinion and inference
34. Use metaphors analogies to express results

Values and Attitudes

1. Appreciate objectivity
2. Be open to new ideas
3. Develop respect for maintenance of natural systems
4. Work as a team member
5. Be aware of careers in science
6. Show preference for scientific method
7. Approach knowledge with skepticism
8. Appreciate the nature of a scientific fact
9. Develop capacity for dispassionate observation
10. Develop capacity for self-discipline
11. Value persistence
12. Excel beyond minimum requirements of a task
13. Develop historical perspective on evolution of biological facts/ideas

Laboratory Skills and Techniques

1. Laboratory glassware
2. Triple beam balance
3. Analytical balance
4. Dissection microscope
5. Compound microscope
6. Electron microscope
7. DNA sequencing
8. Spectrophotometer
9. Centrifuge
10. Autoclave
11. Bacterial staining
12. Histological sectioning
13. Bacterial plating
14. Bacterial tube inoculation
15. Cell culture methods
16. Radioactive material
17. Chromatography
18. Electrophoresis
19. Electronic amplifier
20. Oscilloscope
21. Osmometer
22. pH meter
23. Herbarium
24. Plankton net
25. Computer for data analysis
26. Computerized simulations
27. Titration
28. Scintillation counter
29. Maintenance of a field book
30. Collection of specimens
31. Classification by taxonomic key

Dissection

32. Drawing of an observation
33. Frog
34. Rat
35. Cat
36. Dogfish shark
37. Bony fish
38. Human
39. Mink
40. Fetal pig
41. Dog
42. Rabbit
43. Turtle
44. Flower seeds
45. Vascular plants
46. Nonvascular plants
47. Embryos - animals
48. Clams
49. Sponges
50. Starfish
51. Crayfish
52. Insects
53. Roundworms

We received 136 usable surveys (68 percent return rate) from the random sample, and 40 usable returns (60 percent return rate) from the department chairs of the top 50 biology programs. Organizational variants used for contrasts for the analysis included public vs. private, Ph.D.-granting vs. non-Ph.D.-granting, top 50 vs. random sample, and large departments (upper 50 percent in student enrollment) and small departments (lower 50 percent).

A multivariate discriminant function analysis procedure showed that outcome ratings among department chairs of the random sample varied significantly ($p < .05$) according to whether the program offers the Ph.D. degree, but not according to the size of the department or its public or private status. The results of a second discriminant analysis comparing the importance ratings of Ph.D.-granting programs, large non-Ph.D.-granting programs (student > 45), and small non-Ph.D.-granting programs (students < 45) are presented in table 2. Although there were no significant differences among the groups with respect to cognitive skills, there were significant differences between Ph.D.-granting programs and non-Ph.D.-granting programs with respect to the importance of certain laboratory skills and uses of equipment, and in values and attitudes. Laboratory techniques and equipment usage related to microbiology, vertebrate zoology, invertebrate zoology, and botany were rated significantly ($p < .05$) lower in importance by chairs of Ph.D.-granting programs than by chairs of either the large or small non-Ph.D.-granting programs. Chairs of Ph.D.-granting programs also rated awareness of careers in science and developing the capacity for self-discipline as less important than did chairs of either large or small non-Ph.D.-granting programs.

An implication of these findings is that core outcomes in biology should emphasize cognitive skills to a greater degree than laboratory techniques and equipment usage or values. Cognitive skills objectives appear to be virtually universal. This finding is not surprising, as the outcomes statements in the cognitive domain were developed by research scientists and reflect the component skills of the scientific method of inquiry. In fact, statements 13 through 29 (see table 1), encompassing Inquiry and Problem Solving, are arranged in the sequence of actions one would perform to conduct a scientific study. Thus, one would have to ask how any scientist could fail to rate them as highly important.

Stage 3: Determine generalizable and specific outcomes. An 80/80 criterion was used to determine generalizable outcomes. An outcome was considered as "core" if at least 80 percent of the department chairs rated it as either moderately important or highly important and at least 80 percent of the department chairs indicated the outcome was taught in either the core or the elective curriculum. An outcome was included in the core if it met the 80/80 standard for either the random sample or the top 50. The core set of outcomes consisted of 21 cognitive skills, 16 laboratory techniques or equipment uses, and 6 values or attitudes (see table 3).

We chose not to delineate sets of specific outcomes related to different types of programs. The reasons were that the only significant ($p < .05$) differences in learning outcomes between program types were between Ph.D.-granting and non-Ph.D.-granting programs, and that the domain of important outcomes of the former was actually a subset of the domain of outcomes of the latter. These were differences of degree rather than kind. Instead, another subset of outcomes was identified, which we labeled "Important, but not taught." These were outcomes rated as important by more than 80 percent of the department chairs in either the random sample or the top 50, but taught in fewer than 80 percent of the programs in both the

Table 2.—Discriminant function analysis of importance ratings[a]: Ph.D.-gra
non-Ph.D.-granting programs[c] (n = 59) vs small non-Ph.D.-grant

Domain	Function	Eigen-Values	Canonical Correlation	P
1. Cognitive Skills	1	.444	.55	.78
	2	.149	.36	.99
2. Laboratory Techniques/Equipment Usage	1	1.113	.73	.02
	2	.704	.64	.29
3. Values	1	.204	.41	.04
	2	.136	.35	.18

[a] Importance rating, 1 = not important, 2 = minimally important, 3 = modera
[b] Number of students: median = 180, range 25-550
[c] Number of students > 45
[d] Number of students < 45

gs[a]: Ph.D.-granting programs (n = 17)[b] vs large
n-Ph.D.-granting programs[d] (n = 60)

Significant Discriminating Items (p<.05)				
Item	PhD Programs Mean	Large Programs Mean	Small Programs Mean	P
1. Laboratory glassware	2.9	3.5	3.5	.03
10. Autoclave	2.2	3.0	3.3	.00
11. Bacterial staining	2.5	3.4	3.3	.00
13. Bacterial plating	2.5	3.3	3.3	.00
14. Bacterial tub innoculation	2.3	3.3	3.2	.00
28. Scintillation counter	2.1	2.1	2.6	.03
35. Cat dissection	2.2	3.3	3.2	.00
36. Dogfish shark dissection	2.2	3.0	2.9	.03
43. Turtle dissection	.7	1.8	1.8	.00
44. Flower, seeds dissection	2.5	3.2	3.3	.01
45. Vascular plants dissection	2.5	3.2	3.2	.03
48. Clams dissection	1.7	2.7	2.6	.01
49. Sponges dissection	1.8	2.7	2.6	.01
50. Starfish dissection	1.7	2.8	2.6	.00
51. Crayfish dissection	1.9	2.9	2.7	.00
52. Insect dissection	2.0	2.8	2.7	.02
53. Round worm dissection	1.9	2.7	2.6	.02
5. Be aware of careers in science	2.7	3.5	3.3	.05
10. Develop capacity for self discipline	3.0	3.7	3.5	.01

ant, 3 = moderately important, 4 = very important.

Table 3.—Important outcomes by random sample and top 50 after 80/80 rule[a] was applied

	Outcome	Random sample (n=136)	Top 50 (n=44)
	Cognitive Skills		
	Basic Knowledge		
1.	Specific facts as measured by standardized tests	90/94[b]	(75/95)[c]
2.	Concepts as measured by standardized tests	92/93	90/93
3.	Relationships between concepts	96/96	100/95
5.	Chemical bases of biological phenomena	96/98	100/98
6.	Physical laws related to biological phenomena	91/98	88/93
	Observation and Measurement		
8.	Describe observations appropriately	92/97	95/98
9.	Select appropriate measuring instruments	89/95	(78/90)[c]
10.	Use metric system	88/95	85/85
12.	Determine measurement error	(79/92)[c]	85/90
	Inquiry and Problem Solving		
13.	Demonstrate knowledge of scientific trends	93/93	88/90
15.	Conduct a literature review	93/95	83/85
16.	Propose a research study	81/86	(85/78)[c]
17.	Formulate a hypothesis	82/89	93/85
18.	Design a controlled experiment	81/86	90/80
20.	Conduct lab or field experiment	82/89	(78/75)[c]
22.	Construct graph, table, or chart to express relationships among variables	85/94	93/88
24.	Use statistics to demonstrate relationships between and among variables	80/95	(78/83)[c]
25.	Draw inferences from data set	90/90	93/88
26.	Formulate appropriate generalizations	91/89	95/83
27.	Write a scientific report	93/93	91/98
28.	Give an oral report	80/90	(75/83)[c]

[a] 80% department chair's rated skill as moderately important or very important; 80% department chairs indicated skill taught in core and/or elective curriculum.

[b] % importance/% taught

[c] ratio in parentheses indicates skill did not meet 80/80 criterion.

Table 3.—Important outcomes by random sample and top 50 after 80/80 rule[a] was applied—Continued

Outcome	Random sample (n=136)	Top 50 (n=44)
Laboratory Skills and Techniques		
Skills and Equipment		
1. Laboratory glassware	90/98	(78/90)[c]
3. Analytical balance	84/94	(78/88)[c]
4. Dissection scope	94/99	83/98
5. Compound microscope	96/97	90/95
8. Spectrophotometer	82/93	83/95
9. Centrifuge	85/94	85/98
11. Bacterial staining	84/96	(53/75)[c]
13. Bacterial plating	85/95	(78/93)[c]
14. Bacterial tube inoculation	83/93	(73/88)[c]
17. Chromatography	82/96	(70/88)[c]
18. Electrophoresis	81/89	(75/90)[c]
22. pH meter	93/99	83/95
25. Computer data analysis	91/90	80/80
31. Classification by taxonomic key	82/92	(48/75)[c]
Dissection		
44. Flower seeds	83/93	(53/63)[c]
45. Vascular plants	82/91	(55/70)[c]
Values and Attitudes		
1. Appreciate objectivity	95/91	(95/78)[c]
2. Be open to new ideas	96/91	(98/78)[c]
4. Work as a team member	88/84	(73/58)[c]
6. Show preference for scientific method	89/86	(88/73)[c]
8. Appreciate scientific fact	93/86	93/80
13. Develop historical perspective on evolution of biological facts/ideas	92/83	(90/65)[c]

---

[a] 80% department chair's rated skill as moderately important or very important; 80% department chairs indicated skill taught in core and/or elective curriculum.

[b] % importance/% taught

[c] ratio in parentheses indicates skill did not meet 80/80 criterion.

random sample and the top 50. Three cognitive skills and seven attitudes and values were included in this set (see table 4). Such outcomes may be of interest to biologists who might explore reasons why they are not taught to the degree of their importance.

Stage 4: Identify model indicators. A comprehensive search for possible measures of outcomes in undergraduate biology, conducted nationally and internationally, resulted in the procurement of 14 national tests and 4 foreign tests (2 English, 1 Danish, 1 German) that were related to the core outcomes identified in Stage 3. Upon review (Peterson & Hayward, 1987b), the Graduate Record Examination Practice Test in Biology and the College Board Achievement Test in Biology: Sample Test demonstrated validity (content) as measures of 7 of the core 21 cognitive skills. The English General Certificate of Education (GCE), A-Level Examinations in Biology, administered either by The Associated Examining Board's International Examination or the Oxford and Cambridge School Examination Board, demonstrated ways in which the majority (at least 18 of 21) of the core cognitive skills could be evaluated in a comprehensive national examination with written and practical components.

After the review, we concluded that first, the use of the English tests in the U.S. would be too impractical and costly because of labor-intensive scoring procedures. Second, the exclusive use of our own nationally standardized multiple-choice tests would yield such a limited amount of information concerning either individual student achievement or program quality. Because of these limitations in existing tests, the development of additional measures to encompass the full range of core outcomes is very much warranted.

We thus sought four conceptual frameworks for the formulation of measures along a low-inference to high-inference assessment continuum, and adapted a scheme suggested by the Council on Postsecondary Accreditation (COPA). This scheme includes four ordinal categories denoting levels of confidence in terms of whether an individual student possesses a desired learning outcome. The four levels of inference, from low to high are:

1. **Direct observations.** These are types of indicators derived from actual observance of the skill performance, such as an essay examination, a laboratory practical, a formal scientific report, or an oral presentation. A multiple-choice test is also included in this category, even though the results are more inferential (i.e, the student might have guessed the answer).

2. **Required assessments.** These include the demonstration of cognitive skills and laboratory skills or equipment usage within the context of courses, such as when students construct a graph demonstrating the relationship between variables, write a critique of an article, or use inferential statistics to test the strength of a relationship between variables. In this case, the degree of attainment of an outcome is inferred from the frequency with which students are required to perform skills. The assumption is that the more frequent the requirement, the higher the probability that the student has acquired the knowledge or skill.

3. **Enrichment experiences.** These are the opportunities for the acquisition of knowledge and skills offered in the elective or co-curriculum in which students choose to participate, such as elective seminars, honors theses, or biology clubs. The assumption is

Table 4.—Important outcomes, but not taught[a]

Skills	Importance/ taught random sample[b]	Importance/ taught top 50
Cognitive Skills		
29. Relate findings to scientific theory	71/77	88/75
31. Detect a false claim	75/72	85/60
33. Differentiate fact, opinion, and inference	90/79	90/70
Laboratory Skills and Equipment		
(None)		
Values and Attitudes		
3. Develop respect for maintenance of natural systems[c]	75/93	85/63
5. Be aware of careers in science	90/77	80/45
7. Approach knowledge with skepticism	90/79	90/65
9. Develop capacity for dispassionate observation	93/77	95/60
10. Develop capacity for self-discipline	90/75	88/45
11. Value persistence	90/74	93/58
12. Excel beyond minimum requirements of a task	90/69	95/50

[a] An outcome is included in set if it was rated as important by 80% of chairpersons of either sample as important, and if it was taught in less than 80% of the programs of both samples.

[b] Importance = Percent chairpersons indicating very important or moderately important
Taught = Percent chairpersons indicating skill taught in either core and/or elective curricula

[c] Exception to rule of inclusion

that the attainment of desired learning outcomes is related to the extent of participation in elective experiences.

4. **Proxy Indicators.** These are indicators in which the extent to which students in a program are acquiring core learning outcomes is completely inferred, such as the number of students per microscope in a departmental inventory, the total number of credit hours of chemistry courses required for the major, or the number of faculty seminars open to students held per year. Thus, proxy indicators are variables that may be predictive of learning outcomes.

Indicators for the various cells of the matrix were developed at conferences held at each of the five workgroup institutions. The participants at each conference included the two workgroup members and other invited colleagues. The participants were handed a list of the outcomes (see tables 3 and 4) and were given the directive "Assume we are members of a program review team from the Florida Board of Regents or from the Southern Association of Colleges and Schools (SACS); please tell us how you would demonstrate to us that the skills and values on the list before you are being learned or acquired in your curriculum." Model indicators of core skills and values developed from the conferences are presented in appendix A (cognitive skills), appendix B (laboratory skills and equipment usage), and appendix C (values). Indicators of outcomes that were "Important, but not taught" are presented in appendix D.

Stage 5: Design specifications for model indicators. The indicators contained within the cells of the matrix in Stage 4 can be used to design and develop tests and surveys to gather information pertaining to the attainment of core learning outcomes. The following four tests and instruments could be developed from the model.

1. **Comprehensive mastery test.** Such a test could be administered in the context of required senior seminar, which would include four sections: A multiple-choice test (such as the College Board Achievement Test in Biology); an essay examination modeled after the GCE test in England; a practical examination in which students design and conduct an experiment over the span of several weeks; and an oral presentation of the experiment. Such a comprehensive test may serve as both a capstone integrative learning experience for students and as a direct assessment of student competence.

2. **Faculty survey.** This survey could be administered to each faculty member who teaches an undergraduate course in the curriculum. The faculty members would indicate whether each of the core skills and techniques is evaluated in their courses, and how each is evaluated. Data from the faculty survey can be aggregated to yield the degree to which students are required to demonstrate core skills within the courses of the curriculum. For example, skill x may be assessed in 90 percent of the courses, whereas skill y may be assessed in only 5 percent of the courses.

3. **Student survey.** This instrument can be administered to graduating seniors to assess the extent to which students recall they were required to demonstrate core outcomes in the courses of their programs of study as well as the extent of their participation and involvement in enrichment experiences.

4. **Proxy survey.** An assessment of the quality of teaching and learning across biology programs might be made through an instrument surveying proxy indicators of student achievement. Such an instrument could be completed by department chairs of a regional or national sample of programs. This is a very speculative approach that requires extensive validation before any practical use can be made of it, but it has been attempted--with some success--in the area of foreign language proficiency (Hilton & Grandy, 1986).

Three prototype surveys are appended to the present report: Faculty Survey of Student Learning; Cognitive Skills and Laboratory Equipment; Graduating Senior Survey of Learning in Undergraduate Biology; and the National Survey of Indicators of Student Learning in Undergraduate Biology. A comprehensive mastery test could be developed from the indicators presented in tables 5, 6, and 7 under the column, "Direct Observations."

Caveat

We have not addressed two important issues in the assessment of student outcomes in this report: (1) the degree of proficiency desired of each cognitive skill, and (2) the content domain through which a given skill is demonstrated. These were purposefully omitted because these are the dimensions through which the uniqueness of biology programs are expressed. The faculty of each program must determine the desired mastery levels of each skill, as well as the content domains (e.g., cellular, organismic or environmental) in which those masteries are to be demonstrated. The student and faculty surveys cited above may help a faculty determine the levels of emphasis across skills by revealing the frequency with which the respective skills are assessed across courses and the frequency students are asked to perform them. Based on these data, faculty can engage in a deliberation of whether current assessment practices approach a predetermined desired state. The assumption is that the level of attainment of any skill is a function of the frequency with which it is assessed and the range of content through which it is demonstrated.

Discussion

This paper presented a five-stage process for developing an indicator model of undergraduate student learning in the discipline of biology that can be adapted locally or nationally. The indicator model consists of a two-dimensional matrix of outcome (cognitive skills, laboratory skills, and values) by assessment method (direct to highly inferential). The indicators generated for the cells of the matrix can be used to design and develop mastery tests for the direct assessment of outcomes, as well as for the development of surveys for the indirect assessments of outcomes. These measures can be used both internally or externally as mechanisms for gathering information with which to make decisions about the improvement of teaching and learning practices in biology.

Internal uses:

1. A comprehensive mastery test can be developed to provide an integrative learning experience for students. It could be incorporated in a required senior seminar with a portion of the course grade tied to performance on the test to enhance motivation. A comprehensive

exam may compel students to relate learnings across diverse courses. Depending on how the testing experience is structured (e.g., degree of assistance by an instructor), it could also be used as a program evaluation measure to assess student competence. From the standpoint of student learning, however, such an experience might help students to appreciate science more fully as a method of inquiry into life processes.

2. The faculty and student surveys, used together, may point to shortcomings in the curriculum with respect to the opportunities for acquiring core cognitive skills, laboratory techniques, and values. Such information can be used to identify courses or enrichment activities for development that would enhance the acquisition of skills and values.

3. The student survey may offer the faculty a means for formally assessing the extent to which students are taking part in the enrichment activities offered within the program. Often, students learn to integrate and transfer knowledge and skills acquired in formal courses through activities such as assisting faculty in their research, club projects, visiting lectures, and the like. The transmission of values is also enhanced through enrichment activities (Astin, 1978; Terenzini and Wright, 1987).

4. Administering a comprehensive battery, including a mastery test, a faculty survey, and a student survey, could provide an overview of the ways in which required and elective experiences accumulate and interact to foster the development of core outcomes. Such information would be much more useful for self-study and curriculum development than the administration of student-instructor rating forms within courses. Such forms are tantamount to "happiness indicators" and often do not provide meaningful guidance to faculty for curriculum improvement. A comprehensive battery of instruments would give the faculty a way of answering the question, "Are we providing ample opportunities to learn and are our students acquiring the knowledge, skills, and values we want them to?"

External uses:

1. With the validation of the relationship between certain proxy indicators and scores students earn on a comprehensive achievement battery, programs can be described in terms of characteristics that are empirically associated with student learning. Such proxy indicators may complement or enhance traditional proxy measures of program quality, such as faculty reputation, admission standards, or placement rate in graduate and professional schools. Through the use of multivariate statistics (e.g., canonical correlation, LISREL), a set of proxy indicators could be derived to maximize the prediction of scores on several direct measures of learning.

2. Proxy indicators can be used to describe norms or trends in the discipline from a national perspective. If it can be shown that certain program characteristics are empirically associated with student learning, then the quality of higher education could be monitored on a yearly basis.

3. Learned societies can use proxy measures of student learning to describe programs and to monitor trends in the discipline. Such information can be used to raise questions and issues pertaining to structure of the discipline, the function of undergraduate education in the sciences, and curriculum standards for teaching and learning.

In conclusion, to say that we teach but are not sure what students learn is analogous to saying a salesman sold but wasn't sure the customer bought. Therefore, in order to evaluate the effectiveness of instruction, we must first declare precisely what students should learn and then determine how it can be assessed. This paper offers a paradigm for how this can be accomplished in the discipline of biology. We see no reason why it cannot be adapted for use in other settings or other disciplines.

References

Alverno College, Liberal Learning at Alverno College, 1976.

Astin, A. W. (1978). Four critical years. San Francisco: Jossey Bass.

Bauer, R. A. (1966). Social indicators. Cambridge, MA: MIT Press.

Brown, D. (1970). "A scheme for measuring the output of higher education." In B. Lawrence, G. Weathersby, & V. W. Patterson (eds.), Outputs in Higher Education: Their Identification, Measurement and Evaluation. Boulder, Col.: Western Interstate Commission for Higher Education, 27-40. ED 043-296.

Conrad, C. F. & Blackburn, R. T. (1985). "Correlates of departmental quality in regional colleges and universities." American Educational Research Journal, 22 (2), 279-295.

Dillman, D. A. (1978). Mail and telephone surveys: The total design method. New York: John Wiley.

Gourman, J. (1985). The Gourman Report. A rating of undergraduate programs in American and international universities (5th edition). California.

Hagstrom, W. O. (1971). "Inputs, outputs, and the prestige of university science departments." Sociology of Education, 44, 375-397.

Jones, L. V., Lindzey, G. & Coggeshall, P. E. (1982). An assessment of research-doctorate programs in the United States: Biological Science. Washington, D. C.: National Academy Press.

Jones, D. P. (1985). "Indicators of the condition of higher education." Report prepared for the National Center for Education Statistics.

Kaplan, A. (1963). The conduct of inquiry: Methodology for behavioral science. New York: Harper and Row.

Klopfer, L. (1971). "Evaluation of learning in science." In B. S. Bloom, J. T. Hastings, & G. T. Maclaus (Eds.), Handbook of formative and summative evaluation of student learning. New York: McGraw-Hill.

Land, C. L. (1975). "Social indicator models: An overview." In K. C. Land & S. Spilerman (Eds.), Social indicator models. New York: Russell Sage Foundation.

Landy, F. J. & Farr, J. L. (1983). The measurement of work performance: Methods, theory and applications. New York: Academic Press.

Lehman, A. E. (1987). Peterson's annual guides/undergraduate study: Guide to four-year colleges (17th edition). Princeton, NJ: Peterson's Guides.

Miller, R. I. (1979). The assessment of college performance. San Francisco: Jossey Bass.

National Science Teachers Association (NSTA) (1987, January). Criteria for excellence in biology. NSTA Newsletter, 1 (5).

Oakes, J. (1987). "Conceptual and measurement issues in the construction of school quality." Paper presented at the annual meeting of the American Educational Research Association, Washington D.C.

Peterson, G. W. & Watkins, K. (1979, September). "Identification and assessment of competence: Final report." Florida Competency-based Articulation Project. ERIC ED169 839.

Peterson, G. W. (1982). "A meta-evaluation of a generic skills approach to the evaluation of academic programs." Resources in Education (ERIC), ED 219-398.

Peterson, G. W. & Hayward, P. C. (1987a). "Summative learning in undergraduate biology: A national perspective." Unpublished manuscript. U.S. Department of Education, Office of Educational Research and Improvement (OERI). Contract #400-88-0057.

Peterson, G. W. & Hayward, P. C. (1987b). "A review of measures of summative learning in undergraduate biology." Unpublished manuscript. U.S. Department of Education, Office of Educational Research and Improvement (OERI). Contract #400-87-0057.

Scriven, M. S. (1972). "Pros and cons about goal-free evaluation." Journal of Educational Evaluation, 3 (4), 1-8.

Soloman, L. C. & Taubman, P. J. (1973). Does college matter? New York: Academic Press.

Stufflebeam, D. I., Foley, W. J., Gephart, W. J., Guba, F. G., Hammond, R. I., Merriman, H. O. & Provus, M. M. (1971). Educational evaluation and decision-making. Itasho, Illinois; Peacock Publishers.

Terenzini, P. T. & Wright, T. M. (1987). "Influences on student's academic growth during four years of college." Research in Higher Education, 26(2), 161-179.

Young, D. L., Blackburn, R. T., & Conrad, C. F. (1987). "Research note: Dimensions of program quality in regional universities." American Educational Research Journal, 4 (2), 319-329.

Appendix A

Indicators of Core Cognitive Skills in Biology[a]

Cognitive Skills	Direct Observations (Mastery)	Required Experiences (Offering)	Opportunities (Offerings/Choice)	Proxy Measures (Indirect)
1. Specific facts as measured by standardized tests	1. Modularized national standardized exams tailored to program (e. g. GRE, ETS Achievement Test) 2. Locally developed standardized examinations at course or program level	1. Required number of course credits taken in Biology core and elective curriculum 2. Mean grade point average earned by majors in core and elective courses	1. Percent students participating in national testing program (GRE, MCAT) 2. Average number of course credits taken by majors beyond required core and elective courses	1.The presence of departmental exams at course or program level (yes, no) 2.Required participation in nat'l testing program (yes, no) 3.Required number of credits for core and elective courses as stated in catalogue 4.Total number of undergraduate courses offered in biology as stated in catalogue
2. Concepts as measured by standardized tests	1. Modularized national standardized exams tailored to program (e. g. GRE, ETS Achievement Test) 2. Locally developed standardized examinations at course or program level	1. Required number of course credits taken in core and elective curriculum 2. Mean grade point average earned by majors in core and elective course	1. Percent students participating in national testing program(GRE, MCAT) 2. Average number of course credits taken beyond required core and elective courses	1.The presence of departmental exams at course or program level (yes, no) 2.Required participation in nat'l testing program (yes, no) 3.Required number of credits for core and elective courses as stated in the catalogue 4.Total number of undergraduate courses offered in biology as stated in catalogue

[a] 80% of department chairs in a national random sample rated skill as moderately important or very important; 80% of department chairs indicated skill is taught in core and/or elective curriculum.

Cognitive Skills	Direct Observations (Mastery)	Required Experiences (Offering)	Opportunities (Offerings/Choice)	Proxy Measures (Indirect)
3. Relationships between concepts	1. Essay questions and use of charts or graphs on course examinations 2. Locally developed standardized departmental examinations at course or program level 3. Modularized national standardized exam tailored to program (GRE, ETS Achievement Test)	1. Percent courses in which students construct graphs & charts in reports or examinations 2. Percent courses in which essay questions are used in examinations 3. Junior/senior seminar course required	1. Percent students attending faculty/student seminars/year 2. Percent students electing to write a thesis 3. Percent students taking a junior/senior seminar	1.Required senior thesis (yes, no) 2.Presence of faculty seminars open to students (yes, no) 3.Required junior or senior seminar (yes, no)
5. Chemical bases of biological phenomena	1. ETS Advanced Placement Test in Chemistry 2. Locally developed standardized departmental examination containing items measuring chemical bases of biological phenomena	1. Number of required chemistry courses in major 2. Average grades by majors in chemistry courses 3. Percent final exams in biology with items measuring chemical bases 4. Percent required textbooks with chapters on biochemical bases	1. Percent students taking elective courses in biochemistry	1.Number of chemistry courses required for major as specified in catalogue
6. Physical laws related to biological phenomena	1. ETS Advanced Placement Test in Physics 2. Locally developed standardized departmental examination containing items measuring physical laws related to biology	1. Number of required courses in physics and geology 2. Average grades in courses in physics and geology 3. Percent final exams with items measuring physical laws 4. Percent required textbooks with chapters on physical bases of biological phenomena	1. Percent students taking elective courses in physics and geology	1.Number of physics and geology courses required for majors as specified in catalogue

Cognitive Skills	Direct Observations (Mastery)	Required Experiences (Offering)	Opportunities (Offerings/Choice)	Proxy Measures (Indirect)
8. Describe observations appropriately	1. Essay exam question using previously recorded data from field or lab notes 2. Given a table or a chart with missing data, lines or information, fill in blanks appropriately 3. Find incorrectly stated observations in table or chart with errors 4. Given an object on practical exam, write as many quantitative or qualitative statements about it	1. Percent courses with graded field book or notebook 2. Percent of lab courses with formal drawings of observations (graded) 3. Junior/senior project (thesis) required	1. Percent students electing a thesis or directed individual study (DIS) project 2. Percent students assisting faculty in recording data	1.Number of required field and lab courses as specified in catalogue 2.Junior/senior project (thesis) required as stated in catalogue
9. Select appropriate measuring instrument	1. Given a problem and object of observation, select the appropriate measuring instruments and defend 2. Essay question describing the function of selected measuring instruments	1. Percent lab courses in which students select instrument and defend selection 2. Junior/senior project (thesis) required	1. Percent students taking a DIS (directed individual study) project 2. Percent students electing a junior/ senior project (thesis)	1.Required junior/ senior project (thesis) required as stated in catalogue (yes, no)
10. Use metric system	1. Given a table in English system, make conversion to metric system 2. In lab practical, record observations using metric system	1. Percent final exams using metric observations 2. Percent laboratory reports using metric observations 3. Percent required textbooks and lab texts using metric system	1. Percent students possessing their own metric conversion tables	1.List of equipment in departmental inventory with metric calibrations

Cognitive Skills	Direct Observations (Mastery)	Required Experiences (Offering)	Opportunities (Offerings/Choice)	Proxy Measures (Indirect)
12. Determine measurement error	1. Differentiate accuracy from precision 2. Given a data set, calculate SEM and interpret its meaning	1. Percent lab courses in which experiments require recording multiple observations and calculation of SEM 2. Statistics course required	1. Percent students taking statistics course as elective	1.Statistics course required as stated in catalogue (yes, no) 2.Statistics course listed as an option to fulfill degree requirement
13. Demonstrate knowledge of scientific trends	1. Trace the history of certain biological terms (e.g. cell theory, genetic information transmission, heredity, immunity)	1. History of Biology (or science) course required 2. Percent courses in which exams contain questions about trends 3. Junior/senior special topics seminar is required	1. Percent students electing History of biology (or science) course 2. Percent students taking special topics seminar course as elective 3. Percent students attending a lecture series	1.History of Biology (or science) course required in curriculum 2.Films in library inventory on scientific trends 3.Junior/senior seminar required as stated in catalogue (yes, no) 4.Presence of a lecture series offered by department
15. Conduct a literature review	1. Given a topic, compose an annotated bibliography 2. List several major source guides to scientific literature	1. Percent courses requiring a research report with bibliography 2. A thesis is required for graduation	1. Percent students: (a) electing a junior/senior project (thesis) (b) assisting in faculty research report (c) taking a directed individual study course requiring a report with citations from literature	1.Junior/senior project (thesis) required as stated in catalogue (yes, no) 2.Presence of a reading room with selected journals in biology building (yes, no)

Cognitive Skills	Direct Observations (Mastery)	Required Experiences (Offering)	Opportunities (Offerings/Choice)	Proxy Measures (Indirect)
16. Propose a research study	1. Given a question with a problem statement, propose a research study	1. Percent lab courses requiring students to propose a study 2. A junior/senior research project (thesis) is required	1. Percent students: (a) electing a junior/ senior project (thesis) (b) assisting in faculty research (c) taking a directed individual study course	1 Junior/senior research project (thesis) required as specified in catalogue(yes, no)
17. Formulate a hypothesis	1. Given a question with a problem statement, and a proposed research study, formulate a hypothesis	1. Percent core and elective courses requiring students to formulate hypothesison test or report 2. A junior/senior research project (thesis) is required	1. Percent students (a) electing a junior/ senior project (thesis) (b) assisting in faculty research (c) taking a directed individual study course	1 Junior/senior research project (thesis) is required as stated in catalogue (yes, no)
18. Design a controlled experiment	1. Given a question with a problem and a hypothesis, develop a research design	1. Percent core and elective lab courses requiring students to design a study 2. A junior/senior research project (thesis) is required	1. Percent students (a) electing a junior/ senior project (thesis) (b) assisting in faculty research (c) taking a directed individual study course	1 Junior/senior research project (thesis) is required as stated in catalogue (yes, no)
20. Conduct lab or field experiment	1. Given a question with a research design, describe procedures for conducting a study from beginning through clean-up	1. Percent core and elective courses requiring students to conduct a study 2. A research project or thesis required for graduation	1. Percent students (a) electing a junior/ senior thesis (b) assisting faculty in research (c) taking an internship (d) taking a directed individual study course	1.Research project or thesis specified in catalogue(yes, no) 2.Number and credits of lab courses required as indicated in catalogue 3. Internship required as stated in catalogue

Cognitive Skills	Direct Observations (Mastery)	Required Experiences (Offering)	Opportunities (Offerings/Choice)	Proxy Measures (Indirect)
22. Construct graph, table or chart to express relationships between and among variables	1. Given a data table construct a graph 2. Given a set of observations, construct a table to express relationships	1. Statistics, analytical geometry, or physics course required 2. Percent lab courses requiring a formal lab report with charts and graphs	1. Percent of students presenting visual displays of results of experiment in course or seminar 2.Percent taking statistics, analytic geometry or physics course as elective	1.Statistics, analytical geometry, physics courses required in catalogue (yes, no) 2.Research project or thesis required in catalogue (yes, no)
24. Use statistics to demonstrate relationships between and among variables	1. Given two or more data sets, use descriptive, correlational and inferential statistics to show relationships	1. Statistics or calculus course required 2. Percent lab courses requiring a formal (standard) lab report in which inferential statistics are used	1. Percent of students taking statistics course as an elective	1.Statistics course required in catalogue (yes, no)
25. Draw inferences from a data set	1. Given a research problem, method, and data table, state appropriate inferences or predictions (could be essay or multiple choice test)	1. Percent of lab courses with formal lab reports required 2. Required junior/senior project (thesis)	1. Percent of students (a) assisting in faculty research (b) electing a thesis (c) taking an individual study course	1.Thesis or project required for graduation as stated in catalogue (yes, no)
26. Formulate appropriate generalizations	1. Given a research method and data set, draw appropriate conclusions and generalizations 2. Critique a research article	1. Percent a lab courses requiring formal lab reports 2. Percent courses requiring students to critique article 3. Required junior/senior project (thesis) 4. Comparative course required	1. Percent of students (a) assisting in faculty research (b) electing a senior thesis (c) taking an individual study course (d) taking comparative course as elective	1.Thesis or project required as stated in catalogue (yes, no)

Cognitive Skills	Direct Observations (Mastery)	Required Experiences (Offering)	Opportunities (Offerings/Choice)	Proxy Measures (Indirect)
27. Write scientific report	1. Compose a topical outline of a scientific report and describe the contents typically included in each topical area 2. Submit a scientific report for publication	1. Percent lab courses requiring formal lab reports 2. Required junior/senior project (thesis)	1. Percent of students (a) assisting faculty in writing papers or articles (b) writing research report for departmental compendium of research conducted by students and faculty (c) selecting to write a thesis 2. Percent students taking elective course in scientific writing	1.Thesis required as stated in catalogue 2.Percent student or faculty/student publications 3.Course in scientific writing offered in curriculum
28. Give an oral report	1. List guidelines for giving an oral report 2. Give a presentation of the results of one's own study before a panel of judges	1. Percent courses requiring an oral report 2. Public speaking course is required	Percent of students: (a) making a presentation at Biology Club, seminar, or professional meeting (b) taking a public speaking course as an elective	1.Public speaking course required as stated in catalogue

Appendix B

Indicators of Core Laboratory Skills/Equipment in Biology

Core Lab Skills/Equipment	Direct Assessment	Required Experiences	Opportunities	Proxy Measures
Laboratory glassware (1); Analytical balance (3); Dissection scope (4); Compound microscope (5); Spectrophotometer (8); Centrifuge (9); Bacterial staining (11); Bacterial plating (13); Bacterial tube innoculation (14); Chromatography (17); Electrophoresis (18); pH meter (22); Computer for data analysis (25); Classification by taxonomic key (31); Flower seeds (44); Vascular plants (45)	1.Given an instrument, (a) describe its function (b) describe the mechanical principles (c) list potential error sources (d) describe procedures for proper care 2.Given a lab practical, collect and record observations independently 3.Computer simulation of equipment and observations (e.g. PLATO)	1.Given a faculty question naire, indicate required level of acquaintance with each instrument in each course they teach, (a) observe demonstration (b) perform under direct supervision (c) collect data independently once (d) achieve proficiency; collect data independently 2 or more times	1.Given a student questionnaire, percent graduating seniors achieving the following levels of acquaintance with each instrument, (a) observe demonstration (b) perform under direct supervision (c) collect data independently once (d) achieve proficiency; collect data independently 2 or more times	1.Number of majors per instrument in departmental inventory of instruments

Appendix C

Indicators of Core Values in Biology

Values	Direct Measures	Required Experiences (Offerings)	Opportunities (Offerings/ Choices)	Proxy Measures (Indirect)
1. Appreciate Objectivity	1. Given two reports on a social issue, determine validity of evidence of the two different views 2. Critique an editorial or article in a newspaper or popular magazine	1. Percent courses including a prepared student debate 2. Percent courses requiring students to critique article	1. Percent students attending lectures or seminars on biological issues	1.Existence of departmental lecture series on contemporary issues (yes, no)
2. Be open to new ideas	1. Present arguments on both sides of an issue	1. Percent courses in which guest lecturers are invited 2. Percent courses requiring essay on controversial topic 3. Percent courses including student debates 4. Percent courses discussing nobel laureates	Percent students: 1.Attending visiting lecture series 2.Taking course offerings which include travel to foreign country	1.Average number of publications per faculty, 2.Percent of faculty taking sabbaticals, travel grants each year 3.Existence of departmental guest lecture series (yes, no)
4.Work as a team member	1. Help produce a group research report 2. Help collect community data, but individual interpretation and conclusion 3. State principles of group leadership and group dynamics	1. Percent courses which include a group project with a report	1. Percent students participating in group project sponsored by Biology Club or department 2. Percent students assisting in faculty research 3. Percent students attending informal departmental social activities	1.Presence of an active Biology Club (yes, no)

Values	Direct Measures	Required Experiences (Offerings)	Opportunities (Offerings/ Choices)	Proxy Measures (Indirect)
6. Show preference for scientific method	1. Given a social problem, demonstrate how scientific method can be used to gain an understanding of it 2. Compare scientific method to more intuitive methods for solving a social problem with biological elements	1. Philosophy of science course required	Percent students: 1. participating in Biology Club 2. assisting in faculty research 3. attending lecture series on social /biological issues 4. taking elective courses in philosophy of science	1.Presence of a departmental lecture series on biological bases of social or moral issues (e.g. AIDS) (yes, no) 2.Philosophy of science course required in curriculum (yes, no)
8. Appreciate the nature of a scientific fact	1. Trace the historical development of a scientific fact/idea (e.g. transmission of genetic information) 2. Given an article or editorial, classify statements of fact, opinion and hypotheses	1. Philosophy of Science or History of Biology course required 2. Percent courses in which there is a unit on the "history" of topic 3. Percent courses in which exams include a question on the development of a scientific fact or idea 4. Percent courses discussing contributions of nobel laureates	Percent of students: 1. taking History of biology course as an elective 2. taking elective course in philosophy of science 3. attending lecture series for faculty/students	1.History of Biology or Philosophy of Science required course included in course offerings (yes, no) 2.Philosophy of science course listed in course offerings (yes, no)
13. Develop historical perspective on evolution of biological facts/ideas (see #8 above)				

Appendix D

Indicators of "Important, But Not Taught" Learning Outcomes[a]

Outcomes	Direct Observations (Mastery)	Required Experiences (Offerings)	Opportunities (Offerings/Choices)	Proxy Measures (Indirect)
1. Relate findings to scientific theory (cognitive skill)	1.Given a data set, make a case for whether the data support or contradict the predictions of a theory	1. Percent lab courses requiring formal lab reports 2. Required junior/senior project (thesis)	1.Percent students: (a) assisting in faculty research (b) electing a senior thesis (c) taking an individual study course (d) taking seminar on current topics	1.Thesis or project required as stated in catalogue (yes, no)
2. Detect a false claim (cognitive skill)	1. Given a statement of a scientific fact in a newspaper, develop a research design to verify it. 2. Describe how a theory evolved over time with new evidence. 3. Given a multiple choice test item, state why each alternative answer is either correct or incorrect	1. History of science or biology course required, or philosophy of science course required. 2. Current topics course is required	1. Percent students: (a) attending lecture series on current topics (b) taking a history of science of biology course, or taking a philosophy of science course (c) taking seminar on current topics	1. History of science or biology course required (yes, no) 2. Philosophy of science course required (yes, no)
3. Differentiate fact, opinion and inferences (cognitive skill)	1. Identify statements of fact, opinion and inference in a paragraph(s) and state why. 2. Given an editorial from a newspaper, identify where author is mistaking fact for opinion, fact for inference, opinion for inference.	1. Current topics course required 2. Critical thinking course is required	1. Percent students (a) taking a current topics course (b) taking a critical thinking course	1. Critical thinking course required as stated in catalogue (yes, no)

[a] 80% department heads rated outcome as very important by either random sampler Top 50 or moderately important, but outcome taught in less than 80% of programs in both national random sample and Top-50.

4. Develop respect for maintenance of natural systems (value)	1. Given a proposed change in an ecosystem due to human intervention, identify intended and unintended consequences and give rationale for each	1. Ecology course is required 2. Students required to participate in project to restore or preserve natural systems	1. Percent students participating in sponsored or non-sponsored activity to preserve or restore natural system	1. Ecology course required as stated in catalogue (yes, no) 2. Presence of an active biology club
5. Be aware of careers in science (values)	1. Name occupations in which a BA/BS degree in biology is a necessary qualification	1. Seminar course is required in which there is a unit on careers 2. Approval of program of study by a faculty advisor is required	1. Percent students: (a) attending presentations on careers (b) electing a seminar course containing a unit on career exploration (c) discussing careers with faculty advisor (d) assisting faculty member in research (e) participate actively in biology club	1. Presence of a career information library in department (yes, no) 2. Student/advisor ratio 3. Approval of program of study by faculty advisor required as stated in catalogue (yes, no) 4. Presence of active biology club (yes, no) 5. Percent graduates pursuing careers in biologically related fields
6. Approach knowledge with skepticism (value)	1. Given a statement of a biological fact, describe the history and the conditions under which it is true and false (e.g. genetic information is carried in DNA molecules)	1. Course in History of Science/Biology or Philosophy of Science is required	1. Percent students: (a) electing a course in History of Science/ Biology or Philosophy of Science (b) assisting faculty member in research	1. Course in History of Science, Philosophy of Science is required as stated in catalogue (yes, no)

7. Develop capacity for dispassionate observation (value)	1. Observe an emotionally laden event (e.g. a surgical procedure) and describe the event as fully as possible 2. Given a controversial or emotionally laden problem (e.g. AIDS), analyse its causes, formulate alternative solutiosn and predict their consequences	1. A course in gross human anatomy is required 2. A laboratory course in which students observe and describe live specimens is required	1. Percent students: (a) taking a course in gross human anatomy (b) observing and describing live specimens 2. Average number of field experiences in which students participated in program of study	1. Number of courses indicating field experiences as described in catalogue
8. Develop capacity for self-discipline (value)	1. Design and conduct a biological experiment under the supervision of a faculty member and report findings to an audience	1. Junior/senior project (thesis) required	1. Percent students: (a) electing a junior/senior thesis (b) assisting faculty in research	1. Junior/senior project (thesis) required as stated in catalogue 2. Average number of publications per faculty member per year
9. Value persistence (value) (see #8 above)	1.	1.	1.	1.
10. Excel beyond minimum requirements of a task (value)	1. Voluntarily present the findings of a scientific investigation to an audience either orally or in writing	1.	1. Percent students: (a) participating in biology club, or (b) assisting faculty research voluntarily, or (c) assisting faculty in courses voluntarily, or (d) participating in producing a department publication	1. Number of research articles co-authored by students in a year

A Study of Indicators of College Student Learning in Physics

James S. Terwilliger, J. Woods Halley, and Patricia Heller
University of Minnesota

Our study of undergraduate learning outcomes in physics encompassed three major activities:

1. The development of a table of specifications descriptive of core undergraduate programs for physics majors in 4-year colleges and universities in the United States by:
   a. an analysis of textbooks used in these core programs within a representative sample of schools, and
   b. a detailed study of physics programs in a small number of selected schools.
2. The analysis of the Graduate Record Examination (GRE) Physics Test using both a five-category cognitive classification scheme and a revised scheme for classifying items, and a quantitative comparison of these indicators with commonly employed textbooks and with the table of specifications which we developed.
3. The exploration of alternative indicators derived from computer-based testing systems, as well as paper-and-pencil approaches using special "double" choice-response formats. Our analysis of the costs and benefits associated with these procedures is presented in a later section of this report.

Textbook Survey Results

The textbook survey was planned because there was little or no objective data concerning what is currently taught to physics majors in the United States. To design an indicator, it is first necessary to know what the institutions are attempting to teach. Once that is clear, one can decide whether to design indicators descriptive of the success of different groups of institutions teaching similar curricula or whether to design "normative" indicators to measure the success of institutions in teaching a different (more ideal) curriculum.

To describe the curriculum taught to physics majors in a reasonably quantitative way, we chose to determine what textbooks were used in physics courses taken by physics majors in a sample of colleges in the U.S. Textbooks were selected as a measure of the curriculum taught because they provide a basis for inter-institutional comparison which no listing of course titles or descriptions can. Further, we suspected (and our survey bears out) that a relatively small number of textbooks represents the curriculum taught in the U.S. fairly well. In this procedure, we assume that the textbooks assigned describe what is actually taught. We have no systematic check on this assumption, though it is consistent with qualitative impressions and anecdotal information.

To obtain a sample of schools for the survey, we selected 80 colleges in four categories: highly ranked schools with graduate programs (TG), highly ranked schools without graduate programs (TUG), other schools with graduate programs (RG), and other schools without graduate programs (RUG). Our source for this selection was the American Institute of Physics' (AIP) listing of institutions which offer undergraduate degrees in physics (Ellis, 1986).

To choose the highly ranked (TG) colleges offering graduate programs, we took the top 20 as listed in the 1982 ranking sponsored by the Conference Board of Associate Research Councils (Jones, et.al., 1982). To choose the 20 "other" (RG) colleges offering graduate programs, we made a random selection of 20 schools from the remaining schools offering graduate programs in the AIP list, weighting each school by the number of physics major graduates in 1985. To select the 20 highly ranked schools (TUG) not offering physics graduate programs, we used the list of 50 liberal arts colleges participating in the Second National Conference on "The Future of Science at Liberal Arts Colleges" (Carrier, et.al., 1987) and made a random selection of 20 schools from it, weighting each school by the number of physics major graduates in 1985. (This procedure was forced on us by the fact that no reliable ranking of undergraduate physics programs appears to be available. The development of a reliable indicator of success in undergraduate teaching, such as we propose, can help to meet this need.) Finally, we made a random selection of 20 more colleges not offering graduate programs (RUG) from the AIP list, weighting the schools in the same way.

Each of the 80 schools was sent a letter explaining the project and a form asking for (1) the number of physics majors graduated in the preceding year, (2) a list of courses taken by physics majors and the number of physics majors taking the course in 1985-86, and (3) the textbook used in each course listed. Usable data were obtained from 59 colleges, or about 74 percent of the sample. The number of institutions responding in each category is listed below:

Category of College	Number	Phys. Grads, 1986
TG	14	575
RG	13	222
TUG	19	208
RUG	13	78

The schools returning data graduated 1,083 physics majors in 1986, roughly one-fifth of the total number of graduates in the country. Our sample is skewed toward the elite schools, since we have data on a much larger fraction of the graduates from those institutions. Nevertheless, as shown below, our results do not reveal significant differences between elite and randomly selected colleges. The bias in sampling appears to have little effect on generalizations drawn from the results.

For convenience, we divided the reported curriculum into five categories, closely paralleling other categorizations of curricular material, such as the recommended AAPT curriculum describe later in this report. The categories were: (1) classical and analytical mechanics, (2) electricity and magnetism, optics and waves, (3) statistical mechanics and thermodynamics, (4) quantum mechanics, modern physics and relativity, and (5) electronics, solid state and "other". We did not use data on entry-level physics courses which are also taken by students majoring in many other specialties.

For each college, data on the number of students taught each of the reported texts was entered on a spread sheet which automatically summed the total number of students reported to be taught from that text as well as the fraction of all physics majors taught from the book.

We found that the 20 texts used by 10 percent or more of the students who were physics majors in the schools in the sample accounted for most of the curriculum taught to physics undergraduates after the entry-level course. These texts (indicated in abbreviated notation here), together with the percent of students represented in the sample who use them, are as follows:

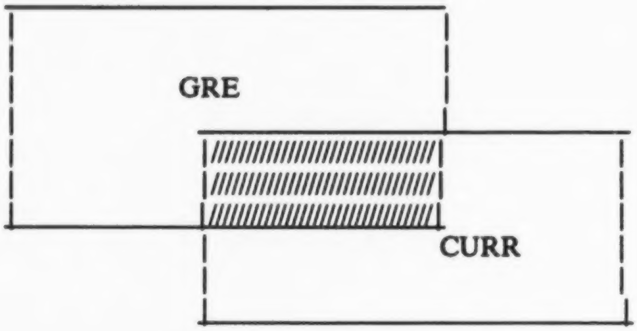
Topics	Text	Percent
Mechanics:		
	Marion	47.7
	Symon	19.9
	Fowles	13.0
	Kleppner	11.1
Electricity and Magnetism:		
	Griffith	23.0
	Hecht & Zajac	18.7
	Purcell	18.1
	Corson & Lorraine	17.4
	Reitz	16.7
Thermodynamics and Statistical Mechanics:		
	Reif	42.6
	Kittel & Kroemer	37.0
Modern Physics:		
	Eisberg	43.0
	Liboff	26.9
	Gasiorowicz	20.2
	Perkins	14.4
	Saxon	13.6
Other:		
	No text	36.7
	Kittel	30.4
	Horowitz	29.1
	Mellisinos	25.0
	Boas	10.4

On the whole, these data confirm our hypotheses that the curriculum taught in the U.S. can be characterized by a relatively small number of textbooks.

Analysis of GRE as a Test of the Curriculum

In the next phase of the study we used the data collected as described above to determine quantitative measures of the extent to which the Graduate Record Examination (GRE) is an appropriate indicator of the success of physics teaching programs for physics majors. The usefulness of such measures is that they can be used first as a general "figure of merit" to decide how well the examination matches the curriculum and then, as we will illustrate, as an analytical tool to determine how the examination or the curriculum could be changed to bring the two into closer congruence.

We imagine the material tested by the GRE to be one set, labelled GRE, and the material taught from the textbooks of the curriculum as another set, labelled CURR (see figure below).



We designed a procedure to estimate the quantity

$$R1 = GRE \cap CURR/GRE$$

and another procedure to estimate the quantity

$$R2 = GRE \cap CURR/CURR$$

R1 may be thought of as the extent to which the curriculum teaches students what they need to know to perform well on the GRE. R2 may be said to be the extent to which the GRE tests the material which is taught from the texts of the curriculum. Put another way, R1 measures the extent to which the curriculum covers the GRE while R2 is the extent to which the GRE covers the curriculum. It should be evident that R1 and R2 are numbers between 0 and 1. In terms of the Venn diagram above, R1 is to be thought of as the ratio of the shaded area to the total labelled GRE, while R2 is the ratio of the shaded area to the total area labelled CURR. If R1 is 1 but R2 is less than 1, it means that everything on the GRE is covered by the curriculum but not everything in the curriculum is tested by the GRE. If R1 is less than 1 but R2 is 1, it means that the GRE tests everything in the curriculum plus some other things not in the curriculum.

We estimate the numbers R1 and R2 as follows. In each case we use an expression of the form:

$$R_i = \frac{1}{Nq(i)} \sum_q \sum_t F(q, t, i)\, f(t)$$

where i = 1 or 2. The index t refers to one of the texts in the curriculum and the sum on t is over all the texts in the sample of the curriculum which we are using. In the results given here, this sample consists of all the texts listed in the preceding section, texts used by 10 percent or more of the students in our survey sample. When we determine Ri for a subfield, the sum on t is restricted to the texts of that subfield as listed in the preceding section. The quantity f(t) is the ratio of the number of students using text t to the total number of students using one or more of the texts in the list included in the sum. We may think of f(t) as an estimate of the probability that the material in text t is taught in the curriculum. (Because some students use more than one text in a given subfield, the sum $\sum_t f(t)$ can be greater than one, even if the sum is restricted to texts within a given subfield.)

The index q refers to questions. When i=1, the sum on q is over all questions on the GRE, or, in the case that the algorithm is applied to a subfield, the sum on q is all questions on the GRE which refer to that subfield. Nq(1) is the number of such questions. For purposes of our analysis we used the sample GRE (Form GR8677, 1985) recently made available to the public by the Educational Testing Service.

The numbers F(q,t,1) are estimates of the extent to which the question q tests the material taught by text t. These numbers were estimated by submitting copies of the sample GRE together with the texts to two experts (physics professors who have taught in the subfield) and asking them to evaluate F(q,t,1) for each question on the GRE in the subfield and for each text. Each professor was thus asked to answer the question, "To what extent would this question q be appropriate (or too easy) for the final examination in a course taught from textbook t?".

The expression

$$\sum_t F(q, t, 1)\, f(t)$$

is approximately the probability that question q is covered by the curriculum so that

$$R_1 = \frac{1}{Nq(1)} \sum_q \sum_t F(q, t, 1)\, f(t)$$

is an estimate of the desired measure. A modicum of error in this estimate arises from the fact that some students use more than one text covering the same material. This effort always leads to over estimates of R1 (and of R2). We have no quantitative assessment of this error, but we do not believe it is larger than ≈ 10 percent of the reported values of Ri.

In the case i=2, the index q refers to questions or problems in the textbook t. A list of twenty such questions was selected at random from each of the textbooks in the curriculum as defined above. The numbers F (q,t,2) are the extent to which a student well prepared to take the GRE would be able to answer the question or do the problem. The quantities F(q,t,2) were again determined by consulting two experienced physics teachers in each subfield, presenting them with lists of twenty problems for each text and a copy of the GRE. Put another way, they answered the question, "If you taught a course to prepare students for the GRE, how appropriate would this question be on a final examination for that course?" The demonstration that he expression for R2 is correct proceeds as in the case i = 1 above.

In table 1 (and figures 1 and 2) we present the R1 results for all five subfields. Results for R2 for all subfields except "Other" are also represented. (We failed to obtain expert judgments in that one case.)

Several effects seem to be significant in these results. For a prominent example, the curriculum in electricity and magnetism is covering, on average, only about 60 percent of the material on the GRE. A study of the detailed data sheets suggests that this phenomenon occurs because there is a significant number of optics questions on the GRE while the standard curriculum does not emphasize that aspect of electricity and magnetism. In fact, only one advanced optics text (Hecht and Zegac) was used by more than 10 percent of our sample. To resolve this particular discrepancy, one could either redesign the indicator to emphasize optics less that does the GRE or encourage institutions to teach more optics. Though there may be an argument in favor of the latter course, we assume that the former is both more desirable and feasible.

Table 1.—R1 and R2 indices for subfield and program type combinations

		TG	TUG	RG	RUG
Electricity and magnetism	R1	0.63	0.58	0.59	0.61
	R2	0.44	0.42	0.42	0.40
Mechanics	R1	0.79	0.76	0.84	0.84
	R2	0.44	0.38	0.48	0.45
Statistical mechanics and Thermodynamics	R1	0.77	0.61	0.82	0.63
	R2	0.17	0.15	0.16	0.18
Modern physics	R1	0.82	0.50	0.85	0.47
	R2	0.24	0.14	0.25	0.14
Other	R1	0.32	0.27	0.28	0.13
	R2				

What then should be added to the assessment to replace the optics questions that are removed? We may get some hint of this from the second line of the table above, which indicates that only 40 percent of the electricity and magnetism curriculum taught in the sample is tested by the GRE. Referring to the questions from texts which the reviewers thought to be particularly unrelated to the kind of questions asked on the GRE, one finds questions emphasizing experimental design and mathematical proofs and derivations related to electricity and magnetism. Again, one may question whether the GRE or the curriculum is encouraging the better kind of learning. If the object is to modify the test to match the curriculum, then one would add questions in which experimental design and mathematical proof and manipulation are emphasized, or weighting the existing questions more heavily in determining a scaled score.

The results on mechanics indicate that the curriculum covers the GRE fairly well but that the GRE is covering only about 40 percent of the curriculum.

With regard to statistical mechanics, the first point is that the number of questions on the GRE in this area is probably too low to reflect correctly the curriculum which is taught. In the sample GRE test with which we worked, there were 10 questions in this area compared to more than 20 in mechanics and in electricity and magnetism, even though as large a fraction of the sample had studied statistical mechanics and thermodynamics. The very low values of R2 for statistical mechanics and thermodynamics (.15 to .18) also indicate that the GRE is not covering this subfield adequately. It is somewhat difficult to be sure of the significance of the discrepancy between the undergraduate and graduate institutions in statistical mechanics which appears in R1 values reported in table 1. But there is no doubt that the discrepancy arises because a smaller fraction of students in undergraduate institutions take statistical mechanics or thermodynamics. This pattern is consistent with anecdotal impressions of faculty who review applications to physics graduate programs.

To add more statistical mechanics and thermodynamics to the assessment instrument would improve the global match between the assessment and the curriculum, but would also result in a bimodal pattern of performance, with students from universities with graduate programs improving their performance relative to students in purely undergraduate institutions. From a normative point of view, this result would pressure the undergraduate colleges to teach more statistical mechanics and thermodynamics (or to require all their majors to take it).

In the case of modern physics, the discrepancy between R1 values for institutions with (.8) and without (.5) graduate programs is even more marked than it is in statistical mechanics and thermodynamics. Again, this is very likely to be because the schools without graduate programs offer or require fewer courses in this area. Modern physics also shows very low values of R2 (.14 to .25) indicating inadequate coverage of this field by the GRE. More questions in this area would increase the bimodalism in performance. Finally, the numbers for the "other" category are not very significant, but they do indicate a poor match between the GRE and the curriculum in the one direction in which we are able to estimate it.

Published Guidelines for Undergraduate Physics Programs

An alternative source of information concerning the undergraduate physics curriculum in U.S. colleges is provided by the Committee on Professional Concerns and Undergraduate Education of the American Association of Physics Teachers (AAPT), AAPT Guidelines for the Review of Baccalaureate Physics Programs (1987).

Figure 1.—Summary of correspondence between the GRE and undergraduate curricula in classical mechanics and electricity and magnetism

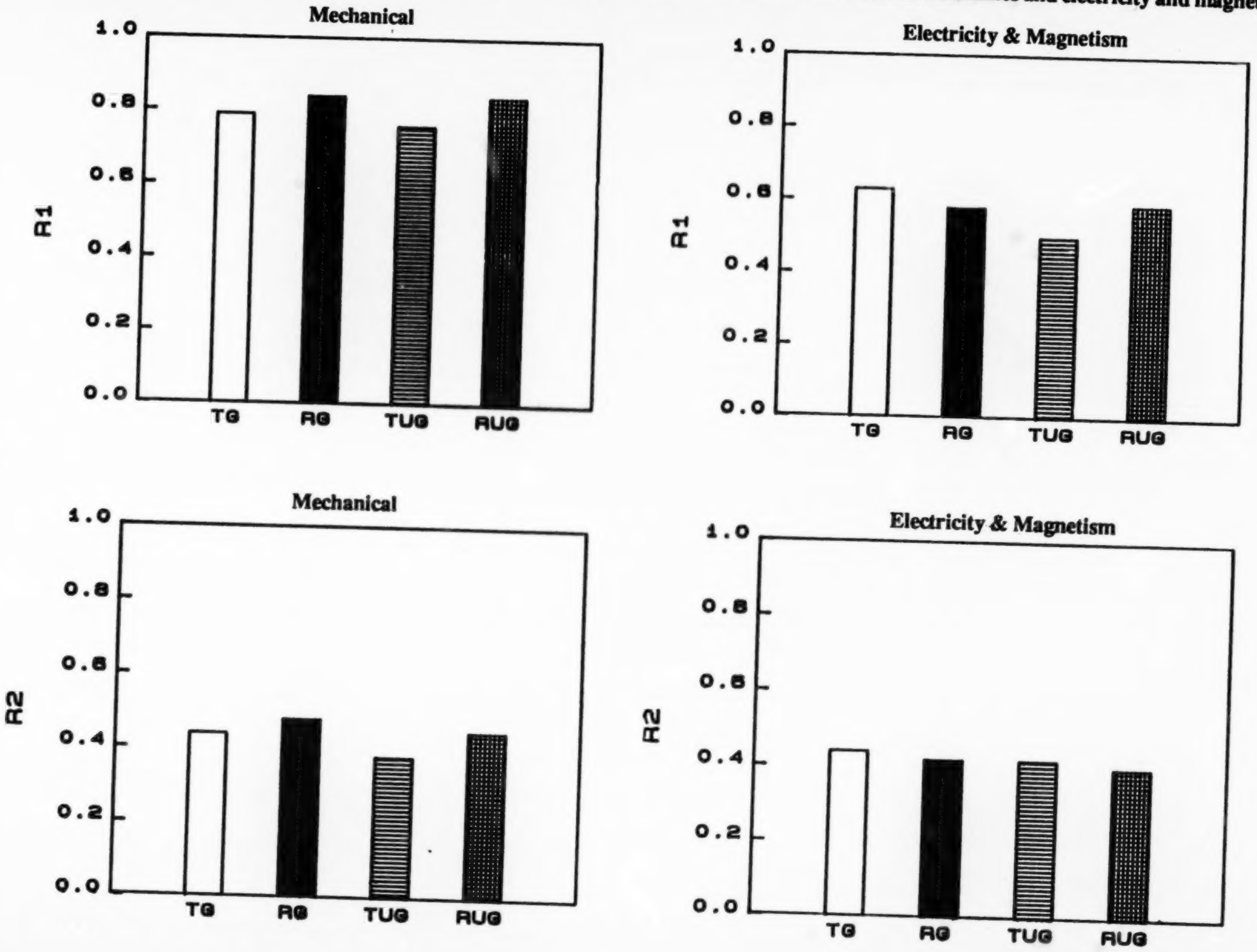
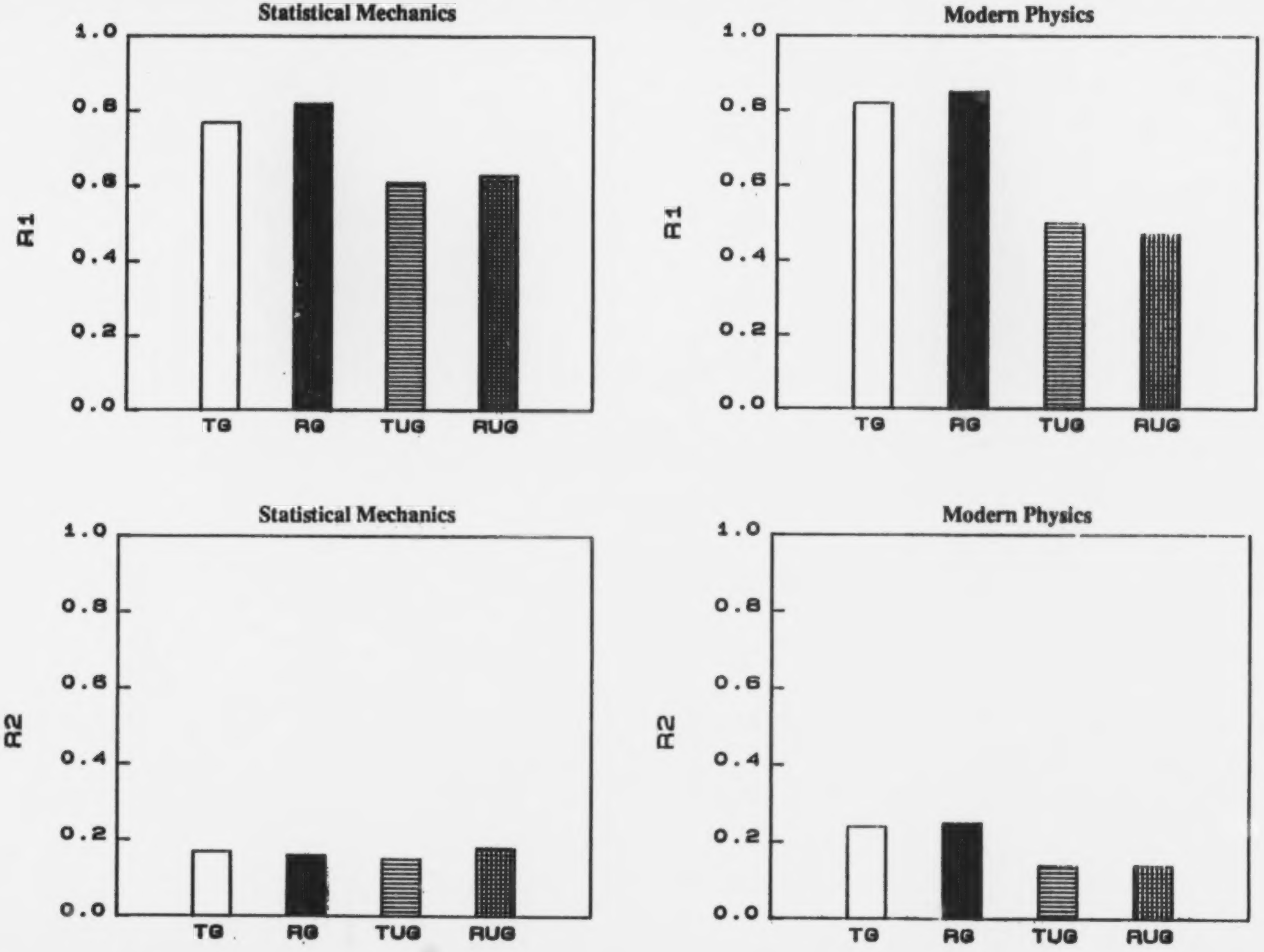


Figure 2.—Summary of correspondence between the GRE and undergraduate curricula in statistical mechanics and modern physics



Statistical Mechanics

Modern Physics

Statistical Mechanics

Modern Physics

The intended uses of the guidelines are described in the introductory section as follows:

> The American Association of Physics Teachers offers these guidelines for the review of baccalaureate physics programs whether at a large comprehensive university or a smaller liberal arts college. We are mindful of the variations this implies concerning curriculum flexibility, staff size and experience, budget support, and physical facilities and equipment. Appropriate comments concerning these variations appear throughout these guidelines.
>
> We emphasize that these are not accreditation standards. These are suggestions for the review of physics programs in departments offering the major in physics. Many of the concerns and questions addressed here are also appropriate for review within departments whose major mission is to provide service courses for other areas. In this case, the reviewer will want to abridge discussion in some areas.

The section entitled "Curriculum" suggests that the undergraduate curricula begin with "an elementary course that has at least five subsections: Mechanics, Waves, Heat and Thermodynamics, Electricity and Magnetism, and Optics. . . . A time commitment of at least two semesters is required to teach the five standard subsections; a time frame of three semesters is a better choice if Modern Physics is to be included."

With respect to more advanced undergraduate courses that a physics major ought to take, the guidelines state, "There should be a vigorous, advanced treatment of topics in Mechanics, Electricity and Magnetism, Thermodynamics and Statistical Mechanics, Optics, Quantum Physics, and Experimental Physics." More specifically, the guidelines present a detailed listing of topics suggested for inclusion in each of these areas. Finally, the guidelines state that the undergraduate curriculum should include a set of elective "capstone" courses, e.g., Astrophysics, Nuclear Physics, Plasma Physics, and Relativity or Solid State Physics. These courses "would normally have as prerequisites one or more of the advanced courses required for a physics major."

The curriculum structure described in the AAPT Guidelines provides an explicit and detailed picture of the course experiences which members of the profession consider to be appropriate for undergraduate majors in physics. In general, our study reveals that the undergraduate curricula at institutions which we surveyed and visited closely parallels that described by the AAPT.

Cognitive Analysis of Outcomes in Physics

Central to our study is the development of a table of specifications for describing "core" undergraduate programs in physics. Such a table has both a "content" and an "outcomes" dimension. The content of physics has been described in some detail in our previous discussions of the curriculum and textbooks. The outcomes of the undergraduate study of physics will be described in this section.

Problem Solving in Physics

Historically, a universally stated goal of instruction in physics is that students become proficient at solving problems. An examination of college-level classroom tests in physics is likely to reveal that most, if not all, questions consist of applied problems that require some combination of factual recall, reasoning, and calculation. Typically, problems posed are quite complex and demand a multistage solution which is constructed by the student. Scoring of responses usually involves a system of "partial credits," which reflects both the quality of the reasoning employed in responding (more important) and the accuracy of calculations performed (less important).

Hewitt (1983) has described the situation from the student's point of view:

> "Students learn very quickly how to play the physics course game. The name of the game is problem solving. I can remember how I played the game when I was a student. Given an assignment (always problem solving), I'd go directly to the back of the chapter and attempt the problems first hand. Some I could do because they were like others I'd done before. To solve the others I'd go back through the chapter looking for the correct formulas. The pages I'd read were the ones with formulas and those with sample problems and solutions. Some pages were all writing--all prose. I had no time for such prose--I had to solve problems! The exams in the course were, of course, all problems. Nothing else. There were no questions calling for qualitative explanations. Some problems required a conceptual understanding of some of the physics, but not all of them. Besides, there was partial credit. What was the value of ideas like the conservation of energy? The value was that in certain problems you could set mgh = 1/2mv2, or in some others, = 1/2 mv2, + Fxs, and in the Fxs part you had to remember to get your sines and cosines right. Unless physics material was a direct aid to solving problems, it had no value." (p. 309)

A significant amount of research on problem solving in physics has appeared in the past decade. Among others, the work of Reif (1981, 1982, 1983); Larkin (1980a, 1980b, 1981); and Chi, Feltovich, and Glaser (1981) has received substantial notice. Much of this literature reflects the strong influence of cognitive psychology and information processing theory.

Cognitive Processes in Problem Solving

A common model for research on problem solving has been to compare and contrast the solutions to physics problems produced by experts and novices. The results have typically revealed that there is little, if any, correspondence between the approaches employed by the two groups. The conceptual framework and reasoning patterns of a student are typically quite different from that of the teacher. This has lead to a greater interest in a more detailed cognitive analysis of students' understanding of basic concepts in physics.

A somewhat more pragmatic concern for greater emphasis on the student's conceptual understanding of physics has been stated by Hewitt (1983) as follows:

> "Let me put it in very strong terms to make my point: A physics student who lacks a conceptual understanding of physics and who is working physics problems is akin to a deaf person writing music or a blind person painting. Too many physics students are cranking away on analytical problems they have no feeling for." (Emphasis in original, p. 309)

Arons (1973) has voiced a similar plea for greater sensitivity to the readiness of students when designing questions and assignments:

> "One of the weakest links in our chain of instruction consists of the questions and exercises that are embalmed at the ends of chapters.... We are desperately in need of collections of questions and problems that, sensitive to the obstacles that arise in students' minds, lead the student through the difficulties and subleties in thinking and reasoning that he must face and overcome. We need questions that challenge his curiosity and ability to perceive relationships but that he can encompass and deal with successfully a reasonable fraction of the time....Above all, we need questions and problems that, gently and gradually, lead the student into extending, inventing, perceiving questions of his own." (p. 781)

Following the lead suggested by Arons, Gray and Lockhead (1980) developed a two-dimensional framework for constructing questions based upon a general cognitive model. The first dimension describes three levels of sophistication required by the content of the question. The second dimension represents three types of action required by the student in response to the question. The result is a nine-cell (three X three) table which attempts to describe all varieties of possible questions from the easiest (student reflexively executes a standard algorithm) to the most difficult (student reflectively constructs a transformation of given information in order to generate new information).

A System for Classifying GRE Items

The two-dimensional model of Gray and Lockhead (1980) was considered to be inappropriate for the GRE because it is quite complex and is designed primarily for free response rather than choice response questions. However, some of the ideas employed by Gray and Lockhead appeared promising as a basis for analysis of GRE items.

The system of categories employed in our analysis is described in figure 3. The system evolved from a series of discussions following ratings of sample items contained in the Physics Test Descriptive Booklet (ETS, 1985). The five categories are defined in an hierarchical fashion with the cognitively least complex represented by category 1 and the cognitively most complex represented by category 6. (Category 3 was not in the original scheme, as we explain below).

To test the system, 25 items were selected at random from Form GR8677 of the GRE Physics Test. These items were given to four judges: two graduate students with extensive undergraduate work in physics and two professors, one in the Department of Physics and the other in Science Education with an MA degree in physics. Each of the four judges independently classified the 25 items using the system shown in figure 3. The results are summarized in table 2.

Figure 3.—Definition of knowledge/skill categories for GRE physics items

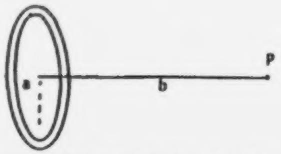
Category	Defining Characteristics	Example
1	Correct response achieved entirely from recall of specific concepts, definitions, equations, etc.	In electrostatic problems, the electric field always satisfies the equation (A) $\nabla \cdot E = \nabla \times E$ (B) $\nabla \cdot E = 0$ (C) $\nabla \times E = 0$ (D) $\nabla (E^2) = 0$ (E) $\nabla(\nabla \cdot E) = \nabla \times E$
2	Correct response achieved from recall of specific equation, principle, or algorithm which is applied in a straight forward manner to a "standard" setting. Highly rehearsed process or procedure	The weight of an object on the Moon is 1/6 of its weight on the Earth. A pendulum clock that ticks once per second on the Earth is taken to the Moon. On the Moon the clock would tick once every (A) 1/6 s (B) $1/\sqrt{6}$ s (C) 1 s (D) $\sqrt{6}$ s (E) 6 s
3	Correct response achieved by analyzing the units of the response options to see which one or two options have the correct units. If two options have the correct units, then additional knowledge may be needed to select the correct option.	 A uniformly charged wire has the form of a circular loop with radius a and a total charge of q. Consider a point P on the axis of the loop a distance b from the loop's center. The potential at point P is (A) $V = (\frac{q}{4\pi\epsilon_o}) \frac{1}{(a^2 + b^2)^{1/2}}$ (B) $V = (\frac{q}{4\pi\epsilon_o}) \frac{ab}{(a^2 + b^2)}$ (C) $V = (\frac{q}{4\pi\epsilon_o}) \frac{1}{b}$ (D) $V = (\frac{q}{4\pi\epsilon_o}) \frac{a}{b}$ (E) $V = (\frac{q}{4\pi\epsilon_o}) \frac{1}{(a^2 + b^2)^{3/2}}$

Figure 3.—Definition of knowledge/skill categories for GRE physics items—Continued

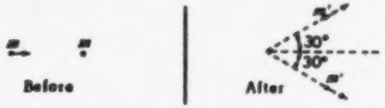
Category	Defining Characteristics	Example
4	Correct response achieved through a qualitative reasoning process based upon the recall of specific concepts, equations, etc. and appropriate application to a "non-standard" setting.	An ideal diatomic gas is initially at temperature T and volume V. The gas is taken through three reversible processes in the following cycle: adiabatic expansion to the volume $2V$; constant volume process to the temperature T; isothermal compression to the original volume V. For the complete cycle described above, which of the following is true? (A) Net thermal energy is transferred from the gas to the surroundings. (B) Net work is done by the gas on the surroundings. (C) The net work done by the gas on the surroundings is zero. (D) The internal energy of the gas increases. (E) The internal energy of the gas decreases.
5	Correct response achieved through a quantitative reasoning process based upon recall of one or more quantitative expressions (or concepts), and appropriate application to a "non-standard" setting. No numerical calculation required although manipulation of formulas may be.	A particle with rest mass m and momentum $mc/2$ collides with a particle of the same rest mass that is initially at rest. After the collision, the original two particles have disappeared. Two other particles, each with rest mass m', are observed to leave the region of the collision at equal angles of 30° with respect to the direction of the original moving particle, as shown below.  What is the speed of the original moving particle? (A) $c/5$ (B) $c/3$ (C) $c/\sqrt{7}$ (D) $c/\sqrt{5}$ (E) $c/2$
6	Correct response achieved through a quantitative reasoning process based upon recall of more than one quantitative expression (or concept) and appropriate application to a given "non-standard" setting requiring a multi-step calculation.	When a narrow beam of monoenergetic electrons impinges on the surface of a single metal crystal at an angle of 30 degrees with the plane of the surface, first-order reflection is observed. If the spacing of the reflecting crystal planes is known from x-ray measurements to be 3 ångstroms, the speed of the electrons is most nearly (A) 1.4×10^{-4} m/s (B) 2.4 m/s (C) 5.0×10^{3} m/s (D) 2.4×10^{6} m/s (E) 4.5×10^{9} m/s

Table 2.—Summary of ratings of 25 GRE items by four judges

A Distribution, by Level of Consensus		B Distribution,* by Model Category		
Level	f	Category	f	$\bar{p}$
4/4	7	5	5	.30
3/4	10	4	2	.38
2/4	7	3	6	.48
1/4	1	2	5	.50
	25	1	3	.86
			21	

*Only 21 items are shown due to the elimination of three items on which there was a 2-2 split among judges and one item on which there was no agreement.

Column A of table 2 shows the inter-judge agreement in terms of the number of judges who place an item in the same category. The level of consensus ranges from complete agreement (4/4) to complete lack of agreement (1/4). As shown in the table, there was agreement as to the cognitive category among at least three of the judges for 17 out of the 25 items. For seven other items, there was agreement among two of the four judges. There was no agreement whatsoever for the remaining item.

Column B of table 2 presents a distribution according to the model category of the items on which two or more judges agreed. All categories are represented and, given the limited number of items, the distribution across categories is roughly uniform. Column B also indicates the mean difficulty level (p) for the items within each category. The increasing progression of p-values as one goes down the column is clearly consistent with the notion that the categories are hierarchically ordered according to the cognitive demands placed upon the examinee. Therefore, the item difficulty data provides support for continued use of a system of cognitive categories.

Summary of Cognitive Analysis

We believe that the level of agreement among judges should be higher than that achieved in table 2, column A. Post-rating discussions among the judges revealed that a major source of disagreement concerned "what a typical examinee knows" at the time of the test administration. To minimize disagreements due to substantial differences between the state of knowledge of judges and typical examinees, we intend to refine the definition of certain categories and use a second panel of judges consisting of seniors and first-year graduate students in physics. These judges will be highly similar to examinees and will be asked to respond in terms of how they actually solve each question.

The issue of the optimal mix of items across the five cognitive categories is admittedly somewhat subjective. Items in category 5 clearly require high-level quantitative ability. However, it is not clear whether the difficulty of such items arises from the reasoning process or the complex calculations which are required. It would seem that less reliance upon "formula sifting" and greater emphasis upon conceptual knowledge could be achieved if more items of the type represented by category 3 were included in place of some items in category 5. This is the direction in which we expect to move in designing an alternative indicator of performance.

A refinement has been made in the original classification system as a result of further analyses. As noted by one of our project consultants, multiple-choice type items provide the respondent with clues which can be used to great advantage by the test-wise examinee. Specifically, he noted that the "correct choice of a formula can be trivially made by a student who understands units, whereas a question requiring a numerical answer requires the student to create a formula (assuming the choices have the same units)." This observation implies that a category which described responses in terms of a "units analysis" of the available options is needed. Such a category was defined resulting in the expanded six-category system shown in figure 3. In the revised system, the new category is inserted (somewhat arbitrarily) between the original second and third categories. Consequently, the three highest (most complex) categories are relabeled as categories 4, 5, and 6.

An attempt was made to test the six-category system by obtaining judgments of GRE items from first-year physics graduate students. Twelve students volunteered to participate, with four students randomly assigned to each of three subfields: classical mechanics, electricity and magnetism, and modern physics. Despite the fact that students were asked to judge only 25 (or fewer) items, only eight students completed the task: three each in classical mechanics and modern physics and two in electricity and magnetism.

Obviously, the sample size is to too small as a basis for meaningful generalizations. However, the results suggest that students generally agree concerning classification of the items in classical mechanics (at least two out of three students classified items the same for 68 percent of the items) and in modern physics (at least two out of three students classified items the same for 78 percent of the items). The two students judging the electricity and magnetism items agree on only 16 percent of their classifications.

The distribution of items across the six categories did not vary greatly from one subfield to the next. In all three cases, the percentage of items classified in the two lowest categories was between 50 and 60, the percentage assigned to the two highest categories was less than 30, and the remaining items were classified in category 4. (Consequently, the addition of category 3 to the system appears to have been in vain.)

Suggestions for Alternative Assessments and Indicators

We have made a preliminary exploration of the possibilities of producing evaluation instruments using microcomputers for administration in place of traditional paper-and-pencil tests.

The possibilities appear to be particularly promising in the following areas:

- o Adaptive testing in fields of achievement already tested by existing instruments such as the GRE;
- o Testing of physics problem-solving skills; and

- o Testing of experimental skills.

We discuss these in turn, confining our attention to issues of feasibility and anticipated problems in implementing a computer-administered test of each type. The development of prototype tests of each type would require an extensive research, development, and evaluation effort.

Adaptive Tests

An adaptive test in physics would be the easiest alternative indicator to produce because extensive work has been done on this aspect of computer administered tests (Weiss, 1985). One can use existing software (for example, see MicroCAT User's Manual, 1987) to construct a tree of test questions. The student is asked questions appropriate to his or her level of expertise, as determined by responses to previous questions. The literature plausibly claims that a more accurate measurement of achievement can be obtained than with standard tests, particularly at the extremes of very high and very low achievement levels.

While this aspect of computer-administered testing is relatively easy to implement, it is unclear whether the resulting modest gains in testing accuracy at the extremes would be worth the time and expense, unless one were particularly interested in detailed information about the performance of students in institutions with very high or very low achievement.

Problem Solving

Our study of textbooks and examinations indicate that a major goal of undergraduate physics education is to teach mathematical problem solving, including writing down the right equations for a given problem, solving the differential equations of mathematical physics, and showing that various equivalent forms of mathematical relations in physics are, in fact, equivalent. None of these skills are adequately tested in the GRE and this is an important reason for the small F values (reflecting expert estimates of curriculum and test overlap) emerging in the study of textbooks described earlier in this report.

The main problem with testing such skills with paper-and-pencil tests at the national level is that the evaluation of performance on the types of tests administered in the classroom is very labor intensive and somewhat subjective. Instructors read the written response of each student with considerable care, because (1) many mathematical forms of response can be equivalent and correct, and (2) students can arrive at the correct response with faulty or incorrect reasoning and mathematical manipulations. Further, it is quite unusual for students to perform perfectly on the multistep questions, so the instructor must track the line of reasoning and mathematical manipulation of the student and then evaluate how far along a correct path the student has gone. This subjective evaluation of student performance is converted by the instructor into a number, the "partial credit" which the student is awarded on the question. The award of partial credit is sufficiently subjective to be the subject of informal negotiation between student and instructor in physics courses. Nevertheless, the process is not as arbitrary as this account may make it appear, as evidenced by the similar grades obtained by a given student on various examinations in similar subjects graded by different instructors and by comparison of the grades awarded by different instructors for the performance of a given student on the same question when such comparisons are made.

The problem of teaching a computer to grade such open-ended examinations has elements in common with many other problems in teaching a computer to perform "expert" functions for which the expert can produce useful and reproducible results without being able to articulate precisely how he or she does it. At the most trivial level, the computer must be able to recognize that many mathematically equivalent forms of response to a question are, in fact, equivalent. Software to perform the needed manipulations is available, for example, in the widely used mathematical manipulation program MAXIMA. This software is not available for microcomputer, but is available (in a form called VAXIMA) for the VAX minicomputer. Thus, writing software to evaluate a final answer to a problem involving mathematical manipulation and an open response format is currently possible. To evaluate "partial credit" would be much harder and we do not see at present how to do it without structuring the questions more than is done on traditional examinations in the physics curriculum. Structured problem solving programs do, however, exist for high school mathematics (e.g., WICAT's High School Geometry Course includes a program that allows students to create geometrical proofs) and for research in physics problem solving (Larkin, et.al. 1980). With sufficient research, development and evaluation, it is possible to produce a computer-administered test that would do a considerably better job of evaluating physics problem-solving skills than the multiple-choice format currently used by the GRE.

Experimental Skills

We found that the GRE, compared to the undergraduate curriculum, does not put as much emphasis on experimental skills. It is not easy to design a paper-and-pencil mass assessment which tests these skills. By experimental skills, we do not mean the ability to manipulate the instruments used in physics experiments. Testing such skills would require actually setting up laboratories for mass testing at a cost which seems likely to be prohibitive on a national scale (although considerable automation of the administration of elementary physics laboratory training has been achieved at some institutions; for example, Michigan Technological University).

Instead, we propose the development of software to test the ability to design experiments. As an elementary example, consider the problem of determining the nature of an electrical device contained in a "black box" with two leads emerging. In the real laboratory environment, the student would be given various meters which he or she could attempt to deduce the electrical nature of the contents of the box. Such an experiment is not difficult to program, perhaps using graphics software like the representation of files used in the MacIntosh operating system. The problem would be in evaluating the response, particularly when the student did not arrive at an entirely correct answer. The problem is similar to the one associated with evaluating responses to mathematical questions but is more difficult because no software for establishing the logical equivalence of various sequences of experiments is available. We conclude that for the present, the use of computer testing to evaluate the ability to design experiments without structuring the situation more than is done in traditional laboratory settings would require some significant development of artificial intelligence software. It may be possible, however, to test for some experimental skills using a more structured simulation which asks a series of questions and offers a limited number of "tries" before the student is shown or asked to interpret the results of a particular test. Such simulations would need to be tested with several experts and novices to devise a reasonable scoring technique for each experimental skill.

The major disadvantages of computer administered tests for experimental skills and mathematical problem-solving ability are

- o the long testing time for the student;
- o the availability of computers at testing sites; and
- o the expense of the research, development, and evaluation of the software.

The physics community would need to agree that the development of such tests is both needed and essential for an adequate (i.e., valid) indicator of undergraduate physics programs at different institutions.

Issues Needing Further Study

This section will review what we believe to be the most important issues associated with the task of developing an alternative indicator of college student learning in physics. While some issues are philosophical, others are more technical in nature.

1. Purpose of the Proposed Indicator

The U.S. Department of Education's RFP to which we responded stated that there is a

> "...dissatisfaction with the type of data available to and used by policymakers and academic leaders to answer the questions 'What are our goals?' 'How are students doing?' and 'What progress are we making?'"

Later in that document we find a call for

> "...the development of 'a concise set of measures,' indicators that would describe 'the "health" of American education' (NCES, 1985). . . .This project is a first step in the development of indicators of that learning, indicators that could be used at many levels, including national, to improve the quality of American higher education."

The language of the RFP is appropriate as a rationale for exploratory steps in the design of indicators, but is not sufficiently specific as a basis for the task of constructing such measures. There is need for more clarity with respect to the intended use of the indicators. Two fundamentally different uses should be considered: program evaluation and student certification.

Program evaluation could be accomplished in large physics departments through some variation of a "matrix sampling" plan in which each student responds to only a sample of all possible items. In programs with relatively few students such an approach would probably not be feasible.

Student certification would require that each student respond to a sufficiently large sample of items to assure a reliable "score" as a basis of judging that individual's competence. Further, if it is desired to certify competence within curricular subfields (classical mechanics, electricity and magnetism, etc.), it would be necessary to design an assessment with a reasonable number of items within each subfield designated.

Clearly, there are important differences in using indicators of program evaluation and indicators of student certification. Further discussion of the issues associated with the purposes of indicators is needed before proceeding with the development of alternative indicators.

2. Content Sampling

There appears to be a broad consensus within the discipline of physics concerning the core areas of the undergraduate curriculum. (See curriculum section on AAPT guidelines.) Our analysis of the GRE suggests that the sampling of content across the core areas in that instrument may not be representative of the emphasis placed upon various content areas in most undergraduate programs. (See, for example, the results of our detailed examination of curricula in the course of visits to six different midwestern colleges and universities in table 3, as well as our previous summary of national data in figures 1 and 2). Specifically, alternative indicators should place somewhat less emphasis upon optics than the GRE and correspondingly greater emphasis upon statistical mechanics, modern physics, and selected laboratory/experimental methods.

Table 3.—Comparison of the content of the GRE in physics with curricula at six institutions*

	Percent of Items	
Content Area	GRE	Curricula
Classical mechanics	18	10
Electricity/magnetism, optics and waves	28	17
Quantum, modern, relativity	32	22
Thermodynamics/statistical mechanics	7	10
Other (includes experimental methods, electronics, and selected advanced topics	15	41
	100	100

*Carleton College, University of Chicago, University of Illinois-Chicago, University of Wisconsin, St. Olaf College, and University of Minnesota.

The design of valid and cost-effective approaches to assessing learning outcomes in laboratory methods will require substantial effort. Traditional paper-and-pencil formats simply do not capture the complexity of "real-world" laboratory settings. It is likely that computer-based exercises will be needed in this area.

The issue of total score vs. subscores, which was previously addressed, has an obvious bearing upon content sampling. If an indicator is designed to reflect student certification at the level of subfields within physics, the concern over relative emphasis across various subfields is less important. (Presumably there will be a sufficiently large number of items in each content area of the assessment to guarantee reliable subscores.) If the indicator employs a total score like the GRE, the issue of representative sampling across content areas seems more significant.

3. Outcome Sampling

As noted earlier in this report, "a universally stated goal of instruction in physics is that students become proficient at 'problem solving'." Evidence for the preeminent status of problem solving is abundant in the classroom physics tests which we collected in visits to the six midwestern departments in the early phases of the project. Problem solving is also prominent in the AAPT guidelines on the undergraduate curriculum.

The cognitive classification scheme which we devised for GRE items (figure 3) explicitly recognizes the importance of problem solving, since three (categories 4-6) of the six categories are defined by problems which require the application of specific qualitative or quantitative reasoning to non standard settings. In other words, these three categories are defined by problem settings which contain novel elements for the typical student. As indicated in table 4, such problems are typically quite difficult (average p-value < .50).

Table 4.—Summary of item difficulties (p) for the GRE physics test for blocks of ten items

		Item Difficulty		
Block	Items	Mean	SD	Range
1	1-10	.50	.16	.19 - .75
2	11-20	.44	.12	.32 - .64
3	21-30	.46	.15	.24 - .71
4	31-40	.58	.25	.20 - .90
5	41-50	.48	.17	.21 - .74
6	51-60	.45	.19	.16 - .79
7	61-70	.51	.21	.22 - .89
8	71-80	.48	.17	.20 - .77
9	81-90	.35	.11	.22 - .57
10	91-100	.28	.10	.14 - .46

A review of approximately 80 tests obtained from over 20 faculty members in physics at the six institutions we visited indicates that classroom test questions are almost exclusively designed to tap outcomes in categories 4-6. Interviews with these faculty members corroborates the fact that the purpose of classroom tests is to assess ability to solve novel problems. This is also consistent with the frequently stated expectation that the "average test score should be approximately 50 percent to 60 percent."

The three remaining categories are defined in terms of conceptual learning (category 1), applications of principles to standard settings familiar to students (category 2), and analysis of response options in search of "correct units" (category 3). The last of these is probably least important since it may be considered as a special strategy employed by "test-wise" students in choice-response situations.

The degree to which an indicator should employ items in categories 1 and 2 is a topic which will require further deliberation. Items which attempt to measure conceptual knowledge in physics are relatively rare in classroom tests and instruments such as the GRE. There is a group of physics educators (a decided minority) who argue that greater emphasis should be placed upon students' conceptual understanding of physics. They claim that physics students frequently are presented with problems for which they lack the prerequisite conceptual knowledge and skills.

To the extent that this viewpoint becomes more widely accepted, the inclusion of more items designed to measure such knowledge may be appropriate. There are several types of paper-and-pencil items that could be used. A simple example (Treagust, 1987) is shown in figure 4. The distracters for the "Reasons" multiple choice items are based on the results of student interviews and student responses to open-ended paper-and-pencil questions designed to identify the most common misconceptions in the content area.

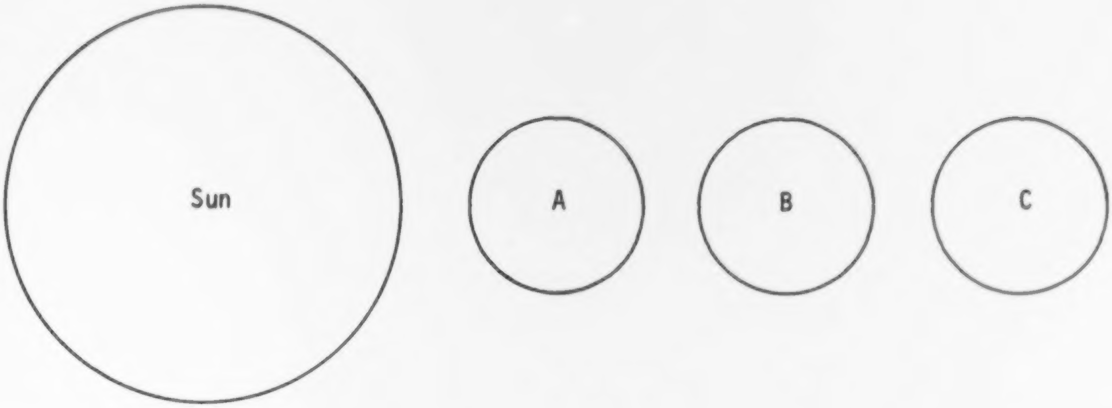
4. Assessment Design and Format

As noted above, problem solving is the primary, if not exclusive, goal of most undergraduate courses in physics. The typical classroom test consists of no more than six problems, each of which is sufficiently complex to require multiple steps (and often several pages) to arrive at an acceptable solution. Points are awarded to responses both on the basis of the quality of the reasoning employed and the accuracy of the calculations (or estimations) performed.

This partial credit approach to scoring free-response questions is appropriate to classroom settings where the number of students is relatively small and/or a teaching assistant is available to share the task of scoring papers. However, a less "labor intensive" approach is needed for assessments designed to provide broader indicators of a student's learning in physics. Specifically, there seems to be no alternative to using a fairly large number of choice-response questions (probably in paper-and-pencil format) if an indicator is to be practically useful on a large scale. In addition, a relatively small set of computer-based exercises might be administered to test selected skills in a more flexible response mode. Such exercises would employ commonly available software and microcomputer systems.

A final question arises as to whether an indicator of student learning based on assessments designed in this format will have "face validity" for physicists. Individuals who are not accustomed to employing choice-response, machine-scorable questions frequently believe that such questions simply cannot measure "important" educational outcomes. It is likely that this view will be common in the discipline of physics. Therefore, it may be necessary to demonstrate empirically the validity of choice-response items through a series of studies employing both free-response and choice-response questions.

Figure 4.—Example questions to test misconceptions



In this solar system, there are three planets the same size. An identical rocket ship is ready to leave each planet. From which planet will it be easiest for the rocket ship to "take off"?

1. Planet A
2. Planet B
3. Planet C
4. The rocket ship will take off from all planets equally easily
5. It is not possible to tell

Reason:

a) Planet C, which is the furthest from the sun, has less gravitational pull from the sun.

b) Planet A, which is closest to the sun, has the highest surface temperature.

c) Planet B, which is neither too hot nor too cold, nor is too close to the sun.

d) Since all of the planets are the same size, the rocketship will take off from all planets as easily.

e) None of the above reasons is correct.

References

Arons, A. (1973). "Toward a wider public understanding of science." American Journal of Physics, 41, 764-781.

Carrier, S.C. and Davis-van Atta, D. (1987). "Maintaining America's Scientific Productivity: The Necessity of the Liberal Arts Colleges," Offices of the Provost and Institutional Research, Oberlin College, Oberlin, OH

Chi, M., Feltovich, P. and Glasser, R. (1981). "Categorization and representation of physics problems by experts and novices." Cognitive Science, 5, 121-152.

Ellis, S.D. (1986). "Data on physics enrollment and degrees," AIP Pub. No. R-121.23A, American Institutes of Physics, NY.

Graduate Record Examination Physics Test, Form GR8677 (1985). Educational Testing Service; Princeton, NJ.

Gray, R. and Lockhead, J. (1980). Can Cognitive Theory Help Us Teach Physics? Technical Report, Cognitive Processes Research Group. Amherst, MA: Department of Physics and Astronomy, University of Massachusetts.

Hewitt, P.G. (1983). "The missing essential--a conceptual understanding of physics." American Journal of Physics, 51, 305-311.

Jones, L.V., Lindzey, G. and Coggeshall, P.E. (1982). "An Assessment of Research-Doctorate Programs in the United States: Mathematical and Physical Sciences," National Academy Press, Washington, DC

Larkin, J., McDermott, J., Simon, D., and Simon, H. (1980a). "Expert and novice performance in solving physics problems." Science, 208, 1335-1342.

Larkin, J., McDermott, J., Simon, D., and Simon, H. (1980b). "Models of competence in solving physics problems." Cognitive Science, 4, 317-345.

Larkin, J. (1981). "Enriching formal knowledge: a model for learning to solve textbook physics problems." In J. Anderson (ed.), Cognitive Skills and Their Acquisition, 311-334. Hillsdale, N.J.: Lawrence Erlbaum.

Reif, F. (1981). "Teaching problem solving: a scientific approach." The Physics Teacher, 19, 310-316.

Reif, F. and Heller, J. (1982). "Knowledge structure and problem solving in physics." Educational Psychologist, 17, 102-127.

Reif, F. (1983). "Understanding and teaching problem solving in physics." In Research on Physics Education: Proceedings of the First International Workshop, 15-53. Paris, France: Centre National de la Recherche Scientifique.

Treagust, D.F. (1987). "An approach for helping students and teachers diagnose misconceptions in specific science content areas." Proceedings of Second International Seminar: Misconceptions and Educational Strategies in Science and Mathematics. Vol. 2, 512-522. Edited by J. D. Novak, Cornell University, Ithaca, NY.

User's Manual for the MicroCAT Testing System (1987). Authored and published by Assessment Systems Corporation: St. Paul, MN.

Weiss, D.J. (1985) "Adaptive testing by computer." Journal of Consulting and Clinical Psychology, 53 (6), 774-789.

Model Indicators of Undergraduate Learning in Chemistry

George M. Bodner
Department of Chemistry
Purdue University

This report describes a model indicator of undergraduate learning in chemistry based in part on the efforts of the American Chemical Society's Committee on Professional Training and the American Chemical Society's (ACS) Examination Institute.

The approach used to construct this model differs from the others described in this volume. Unlike biology or computer science, there was no need to define desired student-learning outcomes--this had been done by the ACS Committee on Professional Training. Unlike the case of mechanical engineering, there was no need to determine clusters of undergraduate experience--clusters representing significantly more than one-half of the undergraduate curriculum had already been identified. Unlike the case of physics, there was no need to solicit data from the departments--5-year reports to the ACS Committee on Professional Training were available for 580 schools offering undergraduate degrees in chemistry.

The archives of the ACS Committee of Professional Training were used as a principal source of information for this project. Five-year reports were studied for 25 percent of the 580 schools that offer ACS-approved degrees in chemistry. About 60 percent of these institutions offer M.S. degrees in chemistry, while about 40 percent offer Ph.D. degrees. This proportion is slightly higher than that the total number of ACS-approved institutions that offer M.S. (54 percent) or Ph.D. (32 percent) degrees. The sample population was divided evenly between private and State-supported institutions. As might be expected, private schools with graduate programs represented a small fraction (25 percent) of the total number of schools offering graduate degrees.

The archives of the ACS Examinations Institute were also used as a source of information for this project. These archives showed how content specialists can construct examinations that can serve as the basis for indicators of undergraduate learning. They also provided information on the development of a comprehensive baccalaureate exam in chemistry that has many of the aspects of a model indicator for summative undergraduate learning in this field.

The ACS Committee On Professional Training

No model for evaluation of undergraduate training in chemistry can be understood without appreciating the role of the ACS Committee on Professional Training in shaping the curriculum. As a result of a survey of unemployed chemists between 1931 and 1934, which suggested that a large proportion of these people were not qualified in training or experience to hold chemical positions, the Committee on Professional Training (CPT) was established in 1936 to study the training of professional chemists and chemical engineers. The first step toward meeting this objective involved analysis of data from a questionnaire sent to chemistry departments. The data solicited included not only chemistry courses, but also essential supporting courses in mathematics, physics, German, English, humanities, and social sciences, as well as instructional and library facilities, departmental administration, staff training, and teaching loads. After summarizing the information from 450 questionnaires, the CPT selected a set of points just above the median of the tabulated data as the minimum standards for the training of undergraduate chemistry majors. Institutions that met these minimum standards were placed on a list of ACS-approved departments. Initially, this list named roughly 100 institutions. In 1956, there were 233 ACS-approved departments; by 1986, this number had increased to 580.

The fact that a department is approved does not mean that all of its graduates meet the minimum requirements for an ACS-approved degree in chemistry; it merely implies that the institution offers work of adequate quality and sufficient quantity to enable a student to meet the minimum standards. Upon graduation, students who meet these standards are certified to the ACS by the head of their department.

CPT Guidelines for Evaluating Undergraduate Education in Chemistry

In summary, an ACS approved program includes

- 400 hours of classroom work in chemistry;
- 500 hours of laboratory work in chemistry;
- a core curriculum that covers the principles of analytical, inorganic, organic, and physical chemistry;
- One year of advanced work in chemistry or allied fields;
- One year of physics; and
- Calculus through differential equations.

Other factors considered in the CPT's evaluation of a program include

- The degree to which the placement of courses is specified. Physical chemistry, for example, shall follow a calculus-based physics course and precede, or at least coincide with, advanced courses in analytical and inorganic chemistry.

- The detail in which the material within and program and the coverage of topics specified.

- The degree of training specified in other areas, e.g., calculus through differential equations (which usually translates into four semesters), and experience with computers that includes programming as well as data acquisition, simulation, information handling and retrieval.

- The emphasis on oral and written communication, which is a result of feedback from the chemical industry that identified these skills as areas in which the training of chemists and chemical engineers has been deficient.

- The attention paid to issues that indirectly influence the quality of undergraduate training, such as teaching loads, equipment and instrumentation, facilities for doing research, nonacademic support staff, and library resources.

- The extent to which the goals of instruction are specified.

- The extent to which the instrumentation and equipment necessary for quality instruction in chemistry is specified.

Further evidence of the role that CPT plays in determining the structure of the undergraduate experience of chemistry majors is provided in appendix A.

The CPT Guidelines as an Indicator of Undergraduate Learning in Chemistry

There was general agreement among the consultants to this project that the CPT guidelines have served as indicators of undergraduate learning in chemistry in the following repects:

- They provide benchmarks that allow individual departments to compare their programs to those offered by similar institutions.

- They specify features of the system known to be linked to desired learning outcomes, such as institutional resources, teaching loads, library facilities, and equipment for research and instruction.

- They define central features of the system, such as the minimum number of chemistry faculty, maximum teaching loads, and minimum equipment and library materials.

- Most importantly, they are generally accepted indicators of minimum standards. On the micro scale, they provide a clear indication of the minimum standards in faculty, facilities, and resources necessary to train chemistry majors. On the macro scale, they trace the rate at which the number of schools offering training in chemistry that meets these standards has increased.

Overlap Of Core Courses Offered To Chemistry Majors

Oakes [1986] argued that indicators should measure features of schooling that are ubiquitous and enduring, and that indicators should be generally accepted as valid and reliable statistics that measure what they are intended to measure. The ease with which an indicator of learning at the undergraduate level can be constructed is proportional to the extent to which the curricula at different institutions overlap. The required chemistry courses for ACS-approved majors at the 146 institutions in this study were therefore analyzed.

All but two institutions offered the equivalent of 1 year of general chemistry: one highly selective institution assumes its students arrive with the equivalent of general chemistry, another selective institution concentrates its offering in general chemistry into a one-semester course. All 146 institutions offer both organic chemistry and physical chemistry. All but two schedule the organic chemistry in the second year; the two exceptions schedule the organic course for the junior year. All offer at least one semester of inorganic chemistry, and a significant number offer two. All but four have a course in instrumental analysis. A significant portion (17 percent) made no explicit mention of a course in quantitative analysis. But many of these schools argued that quantitative analysis had been integrated into their general chemistry course. No program was found that did not require three to four semesters of calculus and 1 year of physics.

Analysis Of Chemistry Textbooks

There is no reason to assume that the textbook defines the curriculum at the undergraduate level the way it does at the primary and secondary level. But it can be argued that the choice of textbook reflects the implied curriculum. One of the first steps toward disclosing the chemistry curriculum, therefore, involved an analysis of the textbooks used at ACS-approved institutions.

Textbooks used in each of the required courses in chemistry were analyzed for the 146 institutions in this study. Some schools failed to provide this information; others listed more than one text for the course. Some schools noted that they use locally-developed materials; others provide information that was obviously erroneous. As a result, the total number of texts cited in a category varied. With the exception of quantitative analysis, the number of texts in each category ranged from 134 to 150. (Analysis of the "quant" course was complicated by the diverse ways in which institutions cover this material.) The most commonly used texts are presented in table 1.

When textbooks used by less than 6 percent of the institutions in this study are ignored, three to six books remain in each category. These books represent the following percentage of the total.

General Chemistry	51.4
Quantitative Analysis	80.0
Organic Chemistry	75.1
Physical Chemistry	82.4
Inorganic Chemistry	86.7
Instrumental Analysis	85.8

This study, therefore, focuses on textbooks used by at least 6 percent of the population.

Table 1.—Analysis of textbook usage

General chemistry	Number of institutions	Percentage of institutions
Brown and LeMay	21	15.0
Brady and Humiston	16	11.4
Masterton and Slowinski	14	10.0
Mortimer	12	8.6
Nebergall, Holtzclaw and Robinson	9	6.4
Quantitative analysis		
Skoog and West	46	46.0
Day and Underwood	15	15.0
Harris	12	12.0
Peters, Hayes and Hiefjte	7	7.0
Organic chemistry		
Morrison and Boyd	39	27.9
Streitweiser and Heathcock	30	21.4
Solomons	25	17.9
Fessenden	11	7.9
Physical chemistry		
Atkins	48	32.4
Alberty and Daniels	16	10.8
Levine	16	10.8
Castellan	15	10.1
Moore	14	9.5
Barrow	13	8.8
Inorganic chemistry		
Huheey	72	48.0
Cotton and Wilkinson	31	20.7
Purcell and Kotz	18	12.0
Jolly	9	6.0
Instrumental analysis		
Skoog	58	43.3
Willard, Merritt and Dean	35	26.1
Christian, et al.,	22	16.4

General Chemistry

Clyde Metz of the College of Charleston collected data for 35 general chemistry textbooks in use at 763 institutions. Analysis of these data demonstrated a remarkable level of agreement among users of each textbook about the sequence in which material should be covered and about what material should be ignored.

Some might argue that this agreement only reflects the extent to which these institutions follow the general outline of the text. The data for 490 institutions were therefore grouped in terms of topics, rather than textbook chapters. The results are reflected in table 2. The first column reports the percentage of institutions which covered this topic in the first semester, while the second column reports similar data for topics in the second semester. (Because of the potential for error when classifying topics in this analysis, data are reported in percentage rounded to one significant figure. And because data in the first two columns must add to 100 percent, entries are only given for the semester in which the largest percentage occurred.) The third column reports the percentage of institutions covering this topic at some point in the two-semester general chemistry course.

Organic Chemistry

Three levels of analysis were used to test the hypothesis that chemistry--particularly organic chemistry--is a high-consensus field (Gage and Berliner, 1984). The first level focused on a list of 100 topics developed in 1985 as a guide for the construction of suitable examination questions for the ACS Cooperative Examination in Organic Chemistry. Some of the subject-matter experts interviewed for this project were tempted to further divide the topics on this list, while others argued with the way topics were organized. But no one questioned whether this material is covered in every full-year organic chemistry textbook.

The topics list cited above was organized from the point of view of a subject-matter specialist in organic chemistry. The second level of analysis examined the list of topics in organic textbooks from the viewpoint of a student encountering organic chemistry for the first time. More than 150 topics were identified that can be found in every one of the textbooks designed for use in the year-long course taken by chemistry majors.

Physical Chemistry

By far the most commonly cited (32 percent) physical chemistry textbook in this study was the text by Atkins. When this text was compared with the others in table 1 that were used for the year-long physical chemistry course for chemistry majors, the order of topics and, to some extent, the depth of coverage was found to vary from one text to another. The extent to which physical chemistry is expressed through equations versus words also varied from one text to another. But similar sets of topics were covered in each text.

Table 2.—General chemistry curriculum organized by topics

	Percent 1st Semester	Percent 2nd Semester	Percent Covering Topic
Units and Measurement	100		100
Atoms, Molecules and Ions	100		100
Elements and Compounds	100		90
Stoichiometry	100		100
Gases	100		100
Electronic Structure of Atom	100		100
Periodicity	100		100
Thermochemistry	100		100
Covalent Bonding	100		100
Molecular Geometry	90		80
Liquids and Solids	80		100
Solutions		60	100
Redox Reactions		80	50
Kinetics		60	100
Qualitative Acid-Base		70	100
Equilibria		90	100
Acid-Base Equilibria		90	100
Ionic Equilibrium		90	80
Thermodynamics		90	90
Electrochemistry		100	90
The Representative Metals		90	50
Nonmetals		80	60
Coordination Complexes		100	60
Nuclear Chemistry		90	70
Organic Chemistry		100	30
Biochemistry		100	10

Inorganic Chemistry

One inorganic chemistry textbook was used by one-half of the institutions in this study; and three texts were used by 80 percent of the population. Instead of analyzing the contents of these texts--as was done for organic and physical chemistry--analysis of the inorganic course focused on the consensus about what topics are important.

The subcommittee responsible for the ACS Cooperative Examination in Inorganic Chemistry distributed a questionnaire to 500 institutions that asked respondents to rate 32 topics in inorganic chemistry as either essential, very important, important, unimportant, or omit. As Hatfield (1979) noted, "The consistency in the results from category to category is remarkable and. . .supports the contention that it is possible to use one examination for a number of purposes, especially if the test items are carefully selected."

The GRE Exams As Potential Indicators

Two measures of summative learning in chemistry already exist--the GRE advanced chemistry exam and the ACS standardized exams. Both appear to satisfy the three criteria for indicators outlined by Jones (1986). They descibe significant aspects of the system, they are policy relevant at both the microscopic and macroscopic levels, and they have been accepted to at least some extent by the primary users. Archival and anecdotal evidence suggests that performance on both measures has been used to assess individual students within a program, to assess the quality of instruction within a chemistry department, and to compare programs.

This project examined evidence for the suitability of these measures to meet two characteristics of an indicator: policy relevance and extent of acceptance. It also probed a question that was not raised explicitly by Jones: "Does the indicator measure a useful quantity?"

As part of this phase, files were examined for 309 students who began graduate work in the Department of Chemistry at Purdue University between 1982 and 1986. Files were also examined for 318 students who applied for admission in 1986 but did not attend Purdue. Data abstracted from the files included the student's undergraduate grade-point average (UGPA), the number of qualifier (or placement) exams passed when the student first enrolled in graduate school, the student's grade-point average at the end of the first year of graduate work, pertinent comments from letters of recommendation that accompanied the application, and percentile rankings of scores on the GRE verbal, quantitative, analytical, and advanced chemistry exams.

A total of 14 hypotheses were tested, including the following:

H1: Admission percentages are high because of preselection of applicants by their undergraduate departments. Analysis of letters of recommendation suggested that undergraduate departments frequently discourage students from applying to Purdue when they believe the student had little or no chance of succeeding in Purdue's program.

H2: Preselection is often based on ACS standardized test scores or GRE scores. Letters of recommendation encouraging admission of a student often noted the student's performance on one of the ACS standardized tests or, less frequently, the GRE exams. Letters that discouraged admission sometimes referred to the student's performance on the ACS tests compared with either students at the institution or national means.

H3: Admission to graduate programs at highly selective institutions is based in part on GRE test scores. The Assessment of Research-Doctorate Programs in the United States: Mathematical and Physical Sciences (Jones, Lindzey, and Coggeshall, 1982) was used to select graduate programs with which Purdue might be compared. Programs for which the mean rating of scholarly quality of program faculty was larger than or equal to 3.0 (n = 47) were contacted to see whether they required GRE scores, particularly the advanced chemistry score, when students applied for admission to graduate school. The departments split more or less evenly between those who required the GRE advanced chemistry exam (n = 16) or strongly recommended it (n = 10) and those that neither required nor recommended the exam (n = 21). Among programs ranked equal to or better than Purdue's, however, more departments either required (n = 9) or recommended (n = 3) the GRE advanced chemistry exam than not (n = 4). Among weaker programs, more did not require the GRE advanced chemistry exam (n = 17) than either required (n = 7) or recommended it (n = 7).

H4: GRE scores are good predictors of success in graduate school in chemistry. Two measures of graduate student performance were used to determine which factors available when a student applies to graduate school best predict success at Purdue: (1) the number of qualifying exams passed when the student first enters Purdue, and (2) the first-year grade-point average. Data collected in a double-blind experiment suggested that students who pass one or more qualifier exam on entrance do better as graduate students than those who do not. The correlation between first-year grade-point average and overall performance as a graduate student is weaker, but still strong. Analysis of these measures provided the following results.

1. A small ($r = .32$), but highly significant ($p < .0007$), correlation was found between the number of qualifiers passed on entrance and the first-year grade-point average.

2. Highly significant correlations were found between the number of qualifiers passed on entrance and the undergraduate grade-point average in science and the GRE verbal, quant, and advanced chemistry exam scores.

    UGPA: $r = .30$ ($n = 111$, $p < .001$)
    verbal: $r = .26$ ($n = 143$, $p < .002$)
    quant: $r = .42$ ($n = 143$, $p < .0001$)
    chem: $r = .68$ ($n = 86$, $p < .0001$)

3. Highly significant correlations were also found between the first-year GPA and the undergraduate GPA in science and the GRE verbal, quant, and advanced chemistry scores.

UGPA: $r = .50$ ($n = 243$, $p < .0001$)
verbal: $r = .25$ ($n = 111$, $p < .000$)
quant: $r = .44$ ($n = 111$, $p < .0001$)
chem: $r = .51$ ($n = 86$, $p < .0001$)

These correlations are consistent with, but much stronger than, those reported in the Guide to the Use of Graduate Record Examinations Program, 1986-87 (ETS, 1986). Correlation coefficients with first-year grade-point averages reported in that document are: UGPA ($r = .34$), verbal ($r = .12$), quant ($r = .23$) and chem ($r = .37$).

4. A stepwise multiple-regression analysis of the factors influencing the number of qualifiers passed on entrance suggested that only one factor was significant when the undergraduate science GPA and GRE verbal, quant, and advanced chemistry exam scores were included in the analysis: the GRE chemistry score.

chem: $R^2 = .456$ ($n = 86$, $p < .0001$)

5. A stepwise multiple-regression analysis of the factors influencing the first-year GPA suggested that two factors were significant when the undergraduate science GPA and GRE verbal, quant, and advanced chemistry exam scores were included in the analysis. The chem score was still the most important factor, but the undergraduate science grade-point average was also significant.

chem: $R^2 = .263$ ($n = 86$, $p < .0001$)
UGPA: $R^2 = .118$ ($n = 86$, $p < .0001$)

H5: The GRE advanced exam in chemistry adequately reflects the content of the undergraduate chemistry curriculum. A panel of subject-matter specialists was asked to judge the extent to which the undergraduate curriculum prepares students for the GRE advanced chemistry exam and the extent to which this exam covers the curriculum (R_1 and R_2 as defined by Terwilliger, Halley, and Heller, 1988). There was some disagreement among members of the panel about the category in which a particular question should be classified. As a result, the questions on this exam were not divided into categories. The value of R_1 obtained in this analysis was 0.85, with a negligibly small standard deviation. The value of R_2 was 0.7, with a significantly larger standard deviation.

These data support the contention that it is possible to construct an indicator of undergraduate learning in chemistry that has predictive power and is therefore likely to be a reliable and valid indicator.

There are well-recognized problems with using the GRE exam for this purpose, however; the most important of which is the self-selected sample of individuals taking the exam. The standardized examinations developed by the ACS Examinations Institute are therefore the basis for the model indicator generated in this project.

The ACS Examinations Institute

The ACS Examinations Committee was created in 1930 with initial funding from the General Education Fund of the Carnegie Foundation. The first test--in general chemistry--was published in 1934. Each test was constructed by the ACS Examinations Committee and then given to the Cooperative Test Service for distribution. CTS also provided technical advice on test construction, a statistical service, and a small subsidy to defray expenses for meetings and correspondence. In 1947, when CTS was merged into the Educational Testing Service (ETS), it became clear that the tests were too expensive for ETS and the ACS Examinations Committee became a self-supporting entity.

The first tests constructed by the ACS Examinations Committee focused on general chemistry--including qualitative and quantitative analysis. Evolution of the ACS Cooperative Examinations program took three different paths. First, there was a gradual move toward including upper-level undergraduate courses, including organic, physical, bio-, and inorganic chemistry. Second, starting in 1957, a series of high-school chemistry exams was developed. Third, graduate-level exams were developed for use by graduate schools to test the backgrounds of their entering students.

The Examinations Committee was recently reorganized as the Examinations Institute, which functions as a unit within the Division of Chemical Education of the American Chemical Society. The Editorial Committee consists of nationally recognized educators from the various fields of chemistry who have been involved with the testing program (typically, as chairs of committees that have constructed tests), and provides expertise and guidance to the currently active test committees. For each test, a chair and a committee of 10 to 40 members actively engaged in teaching that course are appointed by the institute Director.

Features of the ACS Cooperative Examination Program

The following features of the ACS Examinations program should be noted in order to arrive at a judgment of the potential for these examinations to generate both national and institutional indicators of student learning. First, direct comparisons among named institutions are not possible, since the performance of students from a particular department supplying data for national norms is held confidential. Comparisons can be made, however, among institutional types.

Second, the ACS examinations are reliable at the level of the individual student, hence can be used for certification purposes. Given that level of reliability, though, the examination results can be used to make comparisons among different groups of students in the same institution and similar groups of students in different institutions. Given that level of reliability, too, the test results allow instructors to evaluate the strengths and weaknesses in preparation of both individuals and groups, and hence to target instructional efforts more precisely and to advise students more accurately.

The ACS exams have a long history of being used by individual institutions as indicators of undergraduate learning in chemistry. Analysis of the 5-year reports for the sample population in this study provided repeated example of institutions comparing the performance of their students with normative data on the ACS exams. It also provided examples of institutions in which test results on these exams were used to evaluate the strengths and the weaknesses of their students as the basis for making changes in their curricula.

One institution, for example, makes the following arguments for its use of the ACS Examinations programs as comprehensive test for their undergraduate chemistry majors. It argues that these exams provide a reasonable assessment of the student's knowledge base and, in some cases, a comparison of this knowledge base with those of a national population. It argues that these examinations provide an advisement tool that helps identify weak areas so that deficiencies can be eliminated. It notes that recommendations based on these exams have included (1) auditing a course in the area of weakness, (2) independent study and review, (3) extra assignments in existing courses, and (4) personalized instruction. It notes that results on these exams identify weaknesses that would inhibit or prohibit the successful completion of advanced chemistry courses. It also notes that results on these exams can be used to assist students in selecting future goals.

Development of a Senior-Level ACS Baccalaureate Exam

The merit of creating a single senior-level baccalaureate exam (instead of the 20 separate subject-specific exams currently available) has been debated by groups within the American Chemical Society for more than a decade. In 1974, the CPT polled ACS-approved departments to determine their reaction to this suggestion. The cover letter from the CPT in worth quoting at some length, as it confronts many of the sensitive issues in using examination results as the raw data for indicators:

> ". . .there is. . .strong sentiment within the profession for developing some sort of scale against which departments as well as individuals can measure their performance. . . .CPT has held preliminary conversations with representatives of the Educational Testing Service and of the Examinations Committee of the Division of Chemical Education, all of whom felt that the preparation and administration of such an examination would be feasible.
>
> Any plan of this sort would require acceptance and cooperation by the departments involved so we are now turning to you for your reactions and advice. . ."

The letter then went on to describe the source and type of examination.

> "Production of valid examinations is a highly skilled art. At present the Educational Testing Service produces for chemistry the Graduate Record Examination. . . .Within the ACS, the Examinations Committee. . .has produced standardized course examinations for some time and also graduate level examinations in several areas which are widely used for graduate school placement. Other sources may exist. Because of the labor involved and the need for consistency, such examinations are almost always of the objective (true and false or multiple choice) type.

> Presumably any comprehensive achievement examination sponsored by CPT would cover the core program. . .and would include enough questions so that an analysis by area would be statistically valid. In addition advanced sections might be included. We recognize that examinations have their limits. As one CPT member put it, 'If it is to determine the ability of a student to retrieve facts from his. . .head, an examination is possible. If it is to test his ability to interpret given statements or utilize given facts, it may be possible. If it is to test his imagination, creativity, motivation, communication skills, persistence, entrepreneurship, etc., I doubt it.' This limitation must be kept in mind in using the results of any objective examination even though such exams are widely used and correlate with other measures of student academic ability and performance.
>
> We also recognize that a hazard of any uniform examination system is its tendency to freeze programs into a particular pattern. This may be no more serious than the fact that a few textbooks are very widely used. In any event, we believe that the danger would be minimized if exam questions are formulated by a broadly based committee familiar with present teaching practice, and revised frequently. Finally, in order to reduce costs and minimize duplication, it would be highly desirable to use some present examination (perhaps modified in the future by agreement with CPT to make it more compatible with our objectives) rather than to develop an entirely new one. We are currently examining such possibilities."

The cover letter then described the proposed scope of the plan.

> "If an examination program is established, eventually we expect that it would be taken at some time during their senior year by all majors who wish to be 'certified' at schools on the ACS approved list, and would be generally available to anyone else who wished to take it. Initially, we would propose a trial run with a smaller number of departments on a volunteer basis. . . ."

The letter then described potential uses of results.

> "Initially, we propose that the exam results be reported to the individual students and to their departments as national percentiles as well as numerical scores. The students could use the results as part of their record and have them reported to prospective employers or graduate schools, subject to the usual safeguards of academic records. Also, arrangements of some sort would be made for CPT to obtain summaries, aggregated by department, of the results.
>
> . . .if the examination is to establish any validity, it must demonstrate some correlation with professional performance. A comparison with undergraduate grades and grade point average could be made immediately. If a school finds that its students with straight A averages score in the bottom quarter nationally, it should be able to draw some conclusions. However, correlation with subsequent performance, either in professional employment or in graduate school would be more both more significant and more difficult to develop. The ACS would certainly look into the feasibility of gathering and analyzing such data."

The letter then contained a questionnaire, the results of which are summarized in table 3.

Table 3.—Summary of departmental responses to bacculaureate exam questionnaire

1. Should the ACS sponsor a nationally administered achievement examination at the baccalaureate level?

	B*	M*	D*
yes	62	48	72
no	24	11	39
don't know	16	7	7

2. If your answer to 1 is yes, is the proposal outlined in the CPT memo a reasonable approach?

	B	M	D
yes	63	46	70
no	2	4	4

3. Would you be willing to participate in the trial experiment?

	B	M	D
yes	73	46	77
no	26	11	33

4. Should "advanced" sections be included in addition to "core" material?

	B	M	D
yes	29	15	37
no	55	33	50

5. Do most of your majors take a nationally administered exam at present?

	B	M	D
yes	75	39	65
no	24	20	41

(The two most commonly listed exams were GRE and ACS course level exams.)

6. What is your opinion as to the most suitable source for such an examination and its administration?

(Total scores were evenly divided between the present GRE exam, the present ACS exam, a special CPT exam, and a joint CPT-GRE exam)

7. Do you agree with the proposed distribution of results?

	B	M	D
yes	87	48	90
no	12	9	16

8. Would you favor the eventual setting of a passing grade?

	B	M	D
yes	24	20	22
no	68	35	77

9. If such an examination is adopted, should the range and median grades of students entering each Ph.D. program be published in the (ACS) Directory of Graduate Research?

	B	M	D
yes	44	29	24
no	44	26	75

*Institutions whose highest degree offerings are: B = Bachelor's, M = Master's, D = Doctorate.

Almost two-thirds of the departments favored the creation of a senior-level baccalaureate exam. Those that favored the exam were in almost unanimous agreement with the CPT proposal. Slightly more schools were willing to participate in the trial experiment than were in favor of the exam. Only about 40 percent thought that the advanced material should be included, as well as core material. Two-thirds of the departments believed that most of their majors take a nationally administered exam; slightly less than one-half thought they took the GRE; slightly less than one-third reported using the ACS course-level exams. Virtually all of the schools agreed with the proposed distribution of results, but only about one-quarter thought a "passing grade" should eventually be established. Institutions offering only B.S. or B.S. and M.S. degrees were evenly divided about the wisdom of reporting the range and median grades of students entering each Ph.D. program; three-quarters of the Ph.D. institutions were opposed.

Comments, as might be expected, were numerous. Some selected comments are given below.

B.S. Institutions:

> "Status quo. . .allow freedom in teaching. The proposal may lead to licensing of chemists, domination of choice of subjects by the 'in' group, problems about keeping questions confidential, more paperwork. . ."

> "I cannot believe that one exam can be an adequate measure of a student's training and capability as a chemist."

"Isn't ACS accreditation enough? It sure took some doing for a small college of our size to get ACS accreditation. This looks like a new hurdle to me."

"Although it may be better to divide the material into the 'classical' areas of analytical, biochemistry, inorganic, organic, and physical, a more integrated approach to chemistry using divisions such as synthesis, characterizations, dynamics, etc., may well be better."

"I have some reservations because of the tendency of uniform exams to lock us into a pattern. Certainly some provision should be made to study the effect of such an examination system on program development as an integral part of the overall plan."

". . .have mixed feelings. . . .On the positive side, it would allow departments to evaluate and upgrade their programs by comparing their students with others on a nationwide basis. But. . .the ACS course exams give the same information, and we use them extensively in our program as final exams to gain the kind of information that an overall achievement examination would give."

M.S. Institutions:

"The use of standardized exams places a great responsibility, perhaps too great, on the examination or writers of the examination. Diversity and originality are qualities to be prized, but they are increasingly difficult to find in our society."

"Implication that certification might depend on exam scores would destroy certification program--it's hard enough to sell to. . .students as it is!"

"Chemistry is still an experimental science. No written examination has yet been devised to measure lab skills which (judging by recommended lab hours) is still considered important."

"The original goal of CPT was to regulate the training of students--to keep an eye on schools so that students weren't short-changed on their chemical education, that they were exposed to everything they should be. The present goal. . .as read from Gutowsky's letter, is to regulate and judge students, not the training offered by the departments."

Ph.D. Institutions:

"Although we favor the idea of requiring students. . .to take the exam as part of the requirements for certification, we oppose the use of the exam to 'certify' graduates. Poor performance on the exam would obviously be much more to the disadvantage of the individual student than to. . .the institution he attends. We recommend that the aggregated results be used to rate undergraduate departments and to identify those departments which should raise their standards."

"The main value of the ACS accrediting program results when a smallish department is struggling to achieve accreditation. It is a great help to them with persuading their administation to upgrade faculty, library, curricula. Once that has occurred, the benefits are not so great, although the existence of standards do help prevent backsliding. By these means the accrediting program has upgraded U.S. chemistry education very much. However, the classical evil of accrediting programs, especially those depending upon standard examinations is in the great pressures they bring for conformity and uniformity -- in stifling rapid evolution. For example, the N.Y. State Regents' exams (at one time an aid to upgrading high schools) became a disgrace to be eliminated."

As a result of the generally satisfactory response to the CPT questionnaire, a small score experiment was performed in 1975, which involved 113 students from 11 institutions who took the GRE advanced chemistry exam. As might be expected, there was a correlation between the number of chemistry courses the students had taken and their score on the exam.

Number of Courses	Number of Students	Mean Score
6-8	12	535
9-11	35	584
12 or more	41	612

There was a significant correlation between whether they planned to attend graduate school and their score on the GRE exam.

	Number of Students	Mean Score
Will attend graduate school	57	607
Probably will attend graduate school	24	565
Probably will not attend graduate school	5	542
Undecided	2	540

There was a small, but insignificant, difference between the mean score and the highest degree granted by the institution.

	Number of Students	Mean Score
B.S.	55	583
M.S.	34	597
Ph.D.	24	593

There was an equally small difference between the scores of students who had or had not completed an ACS-certified degree program.

	Number of Students	Mean Score
Yes	58	580
No	33	595

The study also found that course grades correlated relatively poorly with subsections of the GRE exam (r = .42 for organic and r = .22 for physical) but better with the overall GRE score (r = .51)

In 1976-77, 1,185 students from 31 schools used an "ACS-Exit Baccalaureate Examination" that was developed by selecting questions from the battery of ACS Cooperative Graduate-Level Placement examinations.

	INORG	ORG	PHYS	ANAL	TOTAL
Number of items on original test	60	75	60	35	230
Number of items used for this test	30	30	20	15	95

Table 4 provides raw scores corresponding to rounded percentiles for the four categories of the test as well as for the total score on this experiment.

Table 4

PERCENT	INORG	ORG	PHYS	ANAL	TOTAL
10	4.3	7.6	4.1	4.6	20.6
20	6.2	9.2	5.1	5.6	26.1
30	8.0	10.3	5.8	6.0	30.1
40	10.0	11.9	6.3	6.5	34.7
50	11.3	13.5	7.1	7.0	38.9
60	13.0	14.8	7.5	7.4	42.7
70	14.9	16.5	8.4	8.1	47.9
80	17.0	18.4	9.3	8.7	53.4
90	20.0	21.0	11.0	9.6	61.6
95	22.6	22.2	12.0	10.3	67.1
100	26.5	26.0	14.7	12.0	79.2

My conclusions after examining the results of the CPT's consideration of a senior-level baccalaureate exam are summarized below.

1. There is general acceptance of the validity of a senior-level exam.

2. The GRE advanced chemistry exam could be used for this purpose.

3. A test constructed by the ACS Examinations Institute would be more likely to reflect the present state of the curriculum.

4. Such a test might also be greeted with less opposition, because the institutions using this test would feel that they have some control its content.

5. There is general agreement with use of a senior-level exam to help departments evaluate their strengths and weaknesses.

6. There is general agreement that normative results from this exam would help departments compare their programs with others.

7. There is some concern about using the results of this exam to evaluate individual students.

8. There is considerable opposition to using the results of this exam to certify individual students.

A Model Indicator Of Undergraduate Learning In Chemistry

The senior-level ACS baccalaureate exam provides an example of how a model indicator of undergraduate learning in chemistry can be constructed. This model indicator has the following advantages:

- It provides an overall score that can indicate how an individual compares with other students within the institution or with a national sample.

- It provides subscores in analytical, inorganic, organic, and physical chemistry, which indicate deficiencies in the background of individual students.

- By providing a pattern of subscores, it compensates for the fact that students are not equally talented in all four of the primary areas of chemistry.

- By providing a pattern of subscores, it helps an institution judge instruction in each of the primary areas of chemistry by comparing the mean performance of their population in that area with their mean performance in the other three.

- It provides a basis for evaluating the effect of changes in instruction both across the curriculum and within one of the primary areas of chemistry.

- It provides a basis by which a department can compare their students with either similar institutions or a national sample.

- It provides information on what a national sample of chemistry majors can (and cannot) do, which will ultimately produce changes in the way chemistry is taught.

- Because all chemistry majors at an institution would take the exam, it eliminates any concern about the sample population.

- Because the exam would be generated by subject-matter specialists working with the ACS Examinations Institute, the content of the exam will be responsive to changes in the curriculum.

-- Because it would be generated by subject-matter specialists who actively teach the course, it is most likely to be accepted by chemists, who will feel they have some control over its content.

-- Because it would be taken at the end of the undergraduate program, the exam would be able to measure the students' retention of information.

The panel of subject-matter specialists who worked with this project argued that there are many limitations to this model indicator.

-- The exam would measure the students' retention of information, but not the students' ability to assimilate new information.

-- Nor would the exam provide an estimate of whether the students are prepared to learn new technical information on their own.

-- Scores on the exam would be more likely to reflect the students' mastery of concepts from the lecture component of the curriculum and not their ability in the laboratory.

-- The exam would not measure creativity.

Members of this panel are now discussing ways in which the ACS exams can be changed to overcome the first three limitations.

References

Gage, N. L., and Berliner, D. C. (1984). Educational Psychology, 3rd Edition, Boston: Houghton-Mifflin.

Grandy, J. (1988). "Indicators of college student learning in the disciplines: Models for assessing achievement in the field of computer science." Final report for the Office of Educational Research and Improvement.

Haenisch, G. and Wiig, G. (1956). The American Chemical Society Committee on Professional Training. Association of American Colleges Bulletin. Volume 42 (2), 321-336.

Hatfield, W. E. (1979). "Inorganic chemistry course survey." Journal of Chemical Education, 54, 359.

Jones, D. P. (1985). "Indicators of the condition of higher education." Report prepared for the National Center for Education Statistics.

Jones, L. V., Lindzey, G., and Coggeshall, P. E. (1982). "An assessment of research-doctorate programs in the United States: Mathematical and physical sciences," Washington, D.C.: National Academy Press.

Oakes, J. (1986). Educational indicators, a guide for policymakers," Center for Policy Research in Education, Occasional Paper OPE-01.

Peterson, G. W., and Hayward, P. C. (1988). "Model indicators of student learning in undergraduate biology." Final report for the Office of Educational Research and Improvement.

Terwilliger, J. S., Halley, J. W., and Heller, P. (1988). "A study of indicators of college student learning in physics." Final report for the Office of Educational Research and Improvement.

Warren, J. (1988). "A model for assessing undergraduate learning in mechanical engineering." Final report for the Office of Educational Research and Improvement.

Appendix A

Appendices to the CPT Guidelines

Appendices to the Guidelines for Evaluation Procedures have been prepared by the ACS Committee on Professional Training for planning courses in the following areas: (1) analytical chemistry, (2) biochemistry, (3) chemical health and safety, (4) chemical information retrieval, (5) computers in chemistry, (6) industrial chemistry, (7) inorganic chemistry, (8) organic chemistry, and (9) physical chemistry. A copy of the appendix specifying the contents of an inorganic chemistry course is given below.

I. Introductory Inorganic Course

- A. Periodicity: basis in atomic structure; classification of the elements; trends in oxidation states, radii, electronegativity; diagonal relationships; term symbols for atomic ground states.
- B. Ionic Interactions: close packed systems; crystal lattices; Born-Haber cycles (lattice energy calculations); ionic radii.
- C. Systematics of the Chemistry of the Elements: alkaline earths, halogens, chalcogens, pnicogens, noble gases, carbon group, boron group, transition elements, lanthanides, actinides.
- D. Acid-Base Chemistry and Non-aqueous Solvents: acid-base concepts, hard and soft acids and bases, non-aqueous solvent systems.

II. Post-introductory Course--Post-First Half of Physical Chemistry

- A. Bonding and Structure: VSEPR theory, symmetry, LCAO-MO theory, valence bond theory, hybridization, bond energies, covalent radii.
- B. Coordination Chemistry: stereochemistry and isomerism, bonding (valence bond, ligand field and MO), magnetic and spectroscopic properties, synthesis, reaction mechanisms, redox chemistry, metal-metal bonds, metal clusters.
- C. Solid State Chemistry: simple metals (structure and bonding), semiconductors, band theory (free electron and tight binding viewpoints), transition metal ions in lattices (spinels), non-stoichiometric solids.
- D. Organometallic Chemistry: EAN rule; carbonyls and nitrosyls; olefin, acetylene, alkyl, arene complexes; metallocenes; oxidative addition and reductive elimination; fluxionality; homogeneous catalysis; organometallic clusters.

III. Below are selected special topics. As many of these topics as possible should be covered in the time available.

- A. Bioinorganic Chemistry: metalloporphyrins, vitamin B12 and cobalamines, nitrogen fixation, metalloenzymes, non-metallic bioinorganic chemistry.
- B. Inorganic Environmental Chemistry.
- C. Boranes.
- D. Inorganic Ring Systems and Polymers: silicates, sulfur nitrides.

IV. Laboratory Course--Concurrent with Post-Introductory Inorganic Course

A selection of syntheses are listed below which demonstrate chemical principles discussed in the above courses. Incorporated into these experiments are a variety of techniques currently used by inorganic chemists and a range of inorganic materials.

Syntheses	Techniques
$K_2S_2O_8$	Electrolysis
Me_3NBF_3	Vacuum manipulation, IR
$NaNH_2$	Non-aqueous solvent
$CrCl_3(THF)_3$	Inert atmosphere box, Soxhlet extractor
$Ni[Pi(OMe)_3]_4$, olefin	Gas chromatography isomerization catalyst
Sm_2OO_3; Nd_2O_3, Pr_6O_{11}	ion exchange
$Co_2(CO)_8$	autoclave, IR
$[Co(NH_3)_5Cl]^{2+}$	UV-visible
$[Co(en)_3]^{3+}$	optical rotations
$[Ni(glycinate)_n]^{(2-n)+}$	pH measurements
$(h^6\text{-}C_6H_6]Cr(CO)_3$	IR, NMR, mass spectra, autoclave
$(h^5\text{-}C_5H_5)_2Fe$ and	TLC and column chromatography derivatives
$[Cr(NH_3)_6](NO_3)_3$	Magnetic susceptibility
$[M_6Cl_{12}]^{n+}$ (M = Ta, Nb)	Dry box technique, high temperature synthesis
ZnS	High temperature transport synthesis
FeF_6^{3-}, $[Fe(CN)_6]^{3-}$	Paramagnetic susceptibility by NMR
$[Re_2Cl_8]^{2-}$	IR, UV-visible

V. Sample Performance Objectives for the core lecture course. (Taken from the 1978 report of the Curriculum Committee of the Division of Chemical Education.)

Periodicity:

1. Define the following with the aid of a suitable example:
   (a) eigenvalue; (b) ionization energy; (c) a node; (d) penetration effects; (c) Zeeman effect; (f) shielding or screening.

2. Indicate the expected variation in properties such as the following as one crosses a given period. (a) covalent radius; (b) electronegativity; (c) ionization energy; (d) electron affinities; (e) natural abundance.

3. Arrange a given set of fluorides (such as InF_3, CF_4, BeF_2, TeF_6, SbF_5) in order of increasing ionic character.

Ionic Interactions:

4. List the thermodynamic quantities necessary to construct a Born-Haber cycle for the formation of an ionic compound from its elements.

5. Given the ionic radii of a cation and anion, calculate their radius ratio and predict the crystalline lattice they will form.

Systematics of the Chemistry of the Elements:

6. Arrange a set of reagents [such as Br_2, $(CN)_2$, F_2, I_2, $(SCN)_2$] in the expected order of oxidizing ability.

7. Discuss the relative stability, molecular formulas, and structure of interhalogen compounds in terms of the nature of the halogens involved.

Acid-Base Chemistry and Non-Aqueous Solvents:

8. Select the stronger acid from pairs such as $B(CH_3)_3$, $BC1_3$; $B(CH_3)_3$, $B(i\text{-}C_3H_7)$; etc.

9. Define the leveling effect.

Bonding and Structure:

10. By way of an M.O. energy level diagram show the interaction of metal d, s, and p orbitals with a set of sigma set of ligand orbitals in an octahedral field.

11. Define the following with the aid of a suitable example: (a) Lewis structure, (b) valence atomic orbital, (c) bonding molecular orbital, (d) antibonding molecular orbital, (e) non-bonding molecular orbital, (f) localized molecular orbital, (g) delocalized molecular orbital, (h) homonuclear, (i) heteronuclear, (j) resonance, (k) lone pair, (l) sigma bond, (m) pi bond, (n) delta bond, (o) Bent's rule, (p) electronegativity, (q) partial ionic character, (r) effective nuclear charge, (s) net bond order, (t) double bond, (u) triple bond, (v) dipole moment, (w) covalent radius, (x) ionic radius, (y) van der Waals radius, (z) sp2 hybrid orbitals, (aa) sp^3 hybrid orbitals, (bb) sp3d hybrid orbitals, (cc) sp^3d^2 hybrid orbitals.

Coordination Chemistry:

12. Stereoisomers:

    a. Given the following complexes, predict the possibility of stereo-isomerism: tetrahedral and square planar MA_2B_2; square planar $M(CD)_2$, where CD represents an unsymmetrical bidentate ligand such as $NH_2CH_2CH_2PH_2$.

    b. Given the octahedral complex $MA_2B_2C_2$, draw structures of all possible stereoisomers. Are these isomers optically active?

Applications of Crystal Field Theory:

13. Given the expected relative magnitude of the radii of metal ions $M^+ : d^1 > d^2 > d^3 > d^4 > d^5$ in the gas phase, explain the "anomalous" order of the observed radii in crystals and complexes, and the "anomalous" order of hydration energies, in terms of crystal field theory.

Synthesis and Reactions:

14. a. Describe the procedure for the synthesis of $[Co(NH_3)_6]^{3+}$ and explain the function of the catalyst which is employed.
    b. Give examples of substitution labile cobalt and platinum complexes.
    c. Give an example reaction for an inner sphere redox reaction and an outer sphere redox reaction.

Solid State Chemistry:

15. Define, in terms of metallic bond theory, insulator, conductor, intrinsic and extrinsic semi-conductor, and p and n-type semi-conductors.

16. List two factors that control the solubility of one metal in another.

17. Compare the spectral, magnetic, and structural properties of the three Robin and Day classes of mixed oxidation state solids.

Organometallic Chemistry:

18. Give an example of the application of Gilman's displacement rule in the preparation of a main-group organometallic.

19. Give the steps in a homogeneous catalysis reaction wherein the catalyst is an organometallic compound.

Bioinorganic Chemistry:

20. List three nonmetals (excluding C, H, O, and N) that are essential to life and briefly indicate one biological function of each.

21. Give an example of the application of the principles of hard and soft acids and bases to the poisoning of enzymes by heavy metal ions and the reversal by agents containing thiol groups (-SH).

Inorganic Environmental Chemistry:

22. Give two examples of the reduction in toxicity of a metal or a ligand via complexation.

Boranes:

23. Give laboratory syntheses for B_2H_6 and BH_4^-.

Inorganic Ring Systems and Polymers:

24. Give laboratory preparations for $(PNC1)_3$, $B_3N_3H_6$, and Na_3O_9.

25. Draw structures for polymers of the silicate, phosphonitrilic, phosphate and sulfur nitride types.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1989— 247-220 / 00626